社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－MPS－73 (23)
2009／3／5

欠損推定量と雑音の分布を用いたマイクロアレイデータ分類

八杉直樹[†] 加藤毅[††,†††] 浅井潔[†,†††]

マイクロアレイはゲノムワイドに遺伝子の発現量を計測する技術で，医療診断など幅広い分野での応用がある．この技術の欠点は，雑音や欠損を含んでしまい，しばしば統計解析による診断結果に悪影響を与える．にもかかわらず，これまでは欠損を補完した値に直接従来の統計解析手法を適用していた．これに対して，本論文ではデータを確率分布とみなして統計解析を行う新しい手法を提案する．実験により，実データを使った提案手法の有効性を示す．

Microarray Data Classification Using Distribution of Imputation and Noise

NAOKI YASUGI,[†] TSUYOSHI KATO[††,†††] and KIYOSHI ASAI[†,†††]

Microarray technique has been used to monitor gene expression levels at a genome wide. A disadvantage of the microarray is that the observed data are error prone; the microarray data often include serious noise and large portions of missing values. Nevertheless, so far the microarray data have been analyzed directly using conventional statistical diagnosis after imputation of missing values. This paper proposes a new automatic diagnostic algorithm which regards the microarray data as a probabilistic distribution. Promising experimental results are shown on real-world data.

1. はじめに

２０世紀末期に細胞の中の遺伝子の発現レベルを測定する装置として，マイクロアレイと呼ばれる技術が開発された．マイクロアレイ技術は１回の実験で，数千から数万の遺伝子の発現量を同時に観測することができる．近年，この技術で得られた遺伝子発現データが細胞生物学，医学，薬学，農学にとって有益な科学的知識を得るために広く用いられており，データの統計解析手法の研究も盛んに行われている[2),3),5)]．例えば，クラスタリングを用いた遺伝子機能推定[2)]，クラス分類手法を用いた疾患分類[3)]や遺伝子ネットワークによる遺伝子制御メカニズムの解析[5)]など数多くの研究があり，さらに可能性は広がっている．

マイクロアレイ技術は，データに雑音や欠損を含んでしまうという問題点がある．雑音や欠損はスライドに傷やほこりがついていることによる環境的要因やスキャナによる画像認識エラーといった技術的な要因など様々な理由から起こる．これら雑音や欠損がしばしば解析結果に悪影響を与える．一つの解決法は，リピート実験を行うことによって，雑音や欠損を軽減する方法であるが，そのための経済的コストは大きい．

従来は，欠損値を何らかの方法で補完してから，データの統計解析が行われてきた．例えば，同じ遺伝子の別の標本の観測値の平均で補完する MEANimpute[6)]が使われており，そのほかにも，KNNimpute[6)]，SVDimpute[6)]，BPCAimpute[4)]，HAImpute[7)]などの欠損補完法が提案されている．多くの場合，これらの方法を使って欠損値を補完してから，従来の統計解析方法を適用していた．しかし，補完された値は推定値に過ぎず，真の値とはずれているはずである．観測された値も雑音がのっているため，真の発現量とはずれていると考えられる．これに対処するためには，ある程度の値の幅を考慮してデータの統計解析を行う必要があるが，これまでは観測されたデータを点としてそのまま解析が行われてきた．

本論文では，マイクロアレイデータの雑音と欠損の問題に対処するため，発現データを点としてではなく，確率分布として扱う新たなクラス分類手法を提案する．本研究では，結果の説明が容易なクラス分類手法として最近傍識別器に着目する．

2. 従来法

本節では，最近傍識別器を使ってマイクロアレイデータを識別するための典型的な手続きを紹介する．

† 東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, University of Tokyo
†† お茶の水女子大学
Ochanomizu University
††† 産総研生命情報工学研究センター
AIST Computational Biology Research Center

遺伝子数を d とし，未知標本を

$$\boldsymbol{x}^{(0)} \equiv [x_{10}, \ldots, x_{d0}]^\top$$

とおく．C 個のクラスからこの未知標本のクラス $y^{(0)} \in \mathbb{N}_C$ を予測したい．ℓ 個の訓練用標本を $(\boldsymbol{x}^{(j)}, y_j) \in \mathbb{R}^d \times \mathbb{N}_C$ を所与とする．記法を簡単にするため，発現データ全体を行列

$$\boldsymbol{X} \equiv \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1\ell} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d0} & x_{d1} & \cdots & x_{d\ell} \end{bmatrix}$$

で表し，訓練用標本のクラスラベルを ℓ 次元ベクトル

$$\boldsymbol{y} \equiv [y_1, \ldots, y_\ell]^\top$$

で表す．なお，行列 $\boldsymbol{X}$ のうち，いくつかの値は欠損している．識別までの流れは次のようになる．

(1) **欠損補完:** 何らかの欠損補完法を使って欠損値を補完する．補完した発現データを

$$\bar{\boldsymbol{X}} = \left[\bar{\boldsymbol{x}}^{(0)}, \bar{\boldsymbol{x}}^{(2)}, \ldots, \bar{\boldsymbol{x}}^{(\ell)}\right]$$

とおく．

(2) **遺伝子選択:** 次元数を減らすため，遺伝子選択を行う．遺伝子選択には，2 クラス分類の場合，t 検定，多クラス分類の場合，F 検定がよく用いられる．選択された遺伝子の添え字の集合を $\mathcal{F}(\bar{\boldsymbol{X}}, \boldsymbol{y})\,(\subseteq \mathbb{N}_d)$ で表すことにする．

(3) **距離計算:** 各訓練用標本と未知標本との二乗ユークリッド距離を計算する．第 j 訓練用標本との二乗ユークリッド距離を $d_j(\bar{\boldsymbol{X}}, \boldsymbol{y})$ とおく．

(4) **投票:** 前のステップで得られた距離の値から距離が近い訓練用標本を探し，その標本のクラスラベルを予測結果とする．K-最近傍識別器を用いる場合，最も距離が近い $K_{\text{classifier}}$ 個を選び，投票により多数決で決める．

このように，従来の方法では，補完した後のデータ $\bar{\boldsymbol{X}}$ を $(\ell+1)d$ 次元のデータ空間における 1 点としてそれ以後のステップを進めている．しかし，実際には真の値とは誤差があるはずである．次節では，欠損推定値の誤差，および観測値の誤差を考慮に入れた新しいクラス分類法を提案するが，その準備のために幾つかの変数を定義する．

距離計算のステップにおける二乗ユークリッド距離は

$$d_j(\bar{\boldsymbol{X}}, \boldsymbol{y}) = \sum_{i \in \mathcal{F}(\bar{\boldsymbol{X}}, \boldsymbol{y})} |\bar{x}_i^{(j)} - \bar{x}_i^{(0)}|^2 \qquad (1)$$

で計算する．ただし，$\bar{x}_i^{(j)}$ は $\bar{\boldsymbol{x}}^{(j)}$ の第 i 成分である．この ℓ 個の距離の値から ℓ 次元ベクトル

$$\boldsymbol{d}(\bar{\boldsymbol{X}}, \boldsymbol{y}) = \left[d_1(\bar{\boldsymbol{X}}, \boldsymbol{y}), \ldots, d_\ell(\bar{\boldsymbol{X}}, \boldsymbol{y})\right]^\top$$

を構成する．

投票のステップにおける $K_{\text{classifier}}$ 個の最近傍の添え字の集合を $\mathcal{I}(\bar{\boldsymbol{X}}, \boldsymbol{y})(\subseteq \mathbb{N}_C)$ で表すとすると，クラス c の獲得票数は

$$s_c(\bar{\boldsymbol{X}}, \boldsymbol{y}) = \left|\left\{i \in \mathcal{I}(\bar{\boldsymbol{X}}, \boldsymbol{y}) \,|\, y_i = c\right\}\right|$$

と表すことができる．各クラスの獲得票数から C 次元のスコアベクトル

$$\boldsymbol{s}(\bar{\boldsymbol{X}}, \boldsymbol{y}) = \left[s_1(\bar{\boldsymbol{X}}, \boldsymbol{y}), \ldots, s_C(\bar{\boldsymbol{X}}, \boldsymbol{y})\right]^\top$$

を構成する．

3. 提 案 法

前節では，従来法が欠損を補完したデータをデータ空間中の 1 点として扱っていることを述べた．本節では，データを確率分布として扱う新たなクラス分類手法を提案する．

簡単のため，データ行列 $\boldsymbol{X}$ の各値は独立に生起すると仮定する．すなわち，データ行列 $\boldsymbol{X}$ の密度関数は

$$p(\boldsymbol{X}) = \prod_{j=0}^{\ell} \prod_{i=1}^{d} p(x_i^{(j)}) \qquad (2)$$

と表せるとする．各発現値は正規分布

$$p(x_i^{(j)}) \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2) \qquad (3)$$

に従うと仮定する．データ $x_i^{(j)}$ が観測されている場合，その正規分布の平均パラメータ μ_{ij} は観測値 $\bar{x}_i^{(j)}$ とし，$x_i^{(j)}$ が欠損値である場合，平均パラメータ μ_{ij} は，欠損補完法によって得られた推定値 $\bar{x}_i^{(j)}$ とする．分散 σ_{ij}^2 の与え方は節 3.1 で述べる．

この分布を使って 3 つの手法を提案する．

手 法 A

一つの方法は，遺伝子選択は欠損補完法から得られる固定した 1 点 $\bar{\boldsymbol{X}}$ を使って求め，距離計算を分布 (2) による期待値で求める方法である：

$$\bar{d}_j = \sum_{i \in \mathcal{F}(\bar{\boldsymbol{X}}, \boldsymbol{y})} \mathcal{E}\left(\left|\boldsymbol{x}_i^{(j)} - \boldsymbol{x}_i^{(0)}\right|^2\right). \qquad (4)$$

この期待値は

$$\begin{aligned} &\mathcal{E}\left(\left|\boldsymbol{x}_i^{(j)} - \boldsymbol{x}_i^{(0)}\right|^2\right) \\ &= \mathcal{E}\left(\left(\boldsymbol{x}_i^{(j)}\right)^2\right) + \mathcal{E}\left(\left(\boldsymbol{x}_i^{(0)}\right)^2\right) - 2\mathcal{E}\left(\boldsymbol{x}_i^{(j)}\right)\mathcal{E}\left(\boldsymbol{x}_i^{(0)}\right) \\ &= (\mu_{ij} - \mu_{i0})^2 + \sigma_{ij}^2 + \sigma_{i0}^2 \end{aligned} \qquad (5)$$

のように閉形式で与えられる．式 (4)，(5) を使って，各標本との距離

$$\bar{\boldsymbol{d}} \equiv \left[\bar{d}_1, \ldots, \bar{d}_\ell\right]^\top$$

を計算し，投票ステップに進む．この手法を**手法 A** と呼ぶ．

手 法 B

手法 A では，遺伝子選択は 1 点 $\bar{\boldsymbol{X}}$ から求めた結果で近似したが，もう一つの距離の計算方法として，そのような近似を行わずに，遺伝子選択も含めた距離の期待値を次のように計算する．各標本との距離のベクトル $\boldsymbol{d}(\boldsymbol{X}, \boldsymbol{y})$ の期待値は

$$\mathcal{E}\left(\boldsymbol{d}(\boldsymbol{X}, \boldsymbol{y})\right) = \mathcal{E}\left(\sum_{i \in \mathcal{F}(\boldsymbol{X}, \boldsymbol{y})} \left|\boldsymbol{x}_i^{(j)} - \boldsymbol{x}_i^{(0)}\right|^2\right)$$

のように表すことができる．第 j 標本との距離の期待値，すなわち，この距離ベクトルの第 j 成分は，期待

値演算の定義より，

$$\mathcal{E}\left(d_j(\boldsymbol{X},\boldsymbol{y})\right)=\int d\boldsymbol{X}p(\boldsymbol{X})\sum_{i\in\mathcal{F}(\boldsymbol{X},\boldsymbol{y})}\left|x_i^{(j)}-x_i^{(0)}\right|^2$$

のように積分で表される．この積分値は，手法 A と異なり，閉形式で与えることができない．そこで，モンテカルロ法を使って近似的に求める．すなわち，分布 (2) にしたがって R 個のデータ行列 $\boldsymbol{X}^{(0)},\ldots,\boldsymbol{X}^{(R)}$ を生成し，距離の期待値を

$$\bar{\boldsymbol{d}}\equiv\mathcal{E}\left(\boldsymbol{d}(\boldsymbol{X},\boldsymbol{y})\right)\approx\frac{1}{R}\sum_{h=1}^{R}\boldsymbol{d}(\boldsymbol{X}^{(h)},\boldsymbol{y})\qquad(6)$$

のように近似的に求める．大数の法則により，R を増やすと漸近的に真の期待値に近づく．このモンテカルロ近似は，生成した各 $\boldsymbol{X}^{(h)}$ に対して，次の計算によって求める：生成した $\boldsymbol{X}^{(h)}$ に対して，遺伝子選択を行い，その結果 $\mathcal{F}(\boldsymbol{X}^{(h)},\boldsymbol{y})$ を式 (1) に代入して $\boldsymbol{d}(\boldsymbol{X}^{(h)},\boldsymbol{y})$ を得る．このようにして得られた距離の平均がモンテカルロ近似である．

ここで述べたモンテカルロ近似によって距離を求める手法を**手法 B** と呼ぶ．

手　法　C

手法 B では距離の期待値を計算したが，獲得票数のスコアベクトルの期待値を計算する方法も考えることができる．スコアベクトルの期待値は

$$\bar{\boldsymbol{s}}\equiv\mathcal{E}\left(\boldsymbol{s}(\boldsymbol{X},\boldsymbol{y})\right)\qquad(7)$$

と表される．この期待値を計算するために必要な積分値

$$\mathcal{E}\left(\boldsymbol{s}(\boldsymbol{X},\boldsymbol{y})\right)=\int dp(\boldsymbol{X})\boldsymbol{s}(\boldsymbol{X},\boldsymbol{y})\qquad(8)$$

も閉形式で与えられないので，モンテカルロ近似で求める：

$$\bar{\boldsymbol{s}}\approx\frac{1}{R}\sum_{h=1}^{R}\boldsymbol{s}(\boldsymbol{X}^{(h)},\boldsymbol{y}).\qquad(9)$$

ただし，R はモンテカルロ近似におけるサンプリング数である．計算方法は次のようになる：各 $\boldsymbol{X}^h$ に対して，遺伝子選択 $\mathcal{F}(\boldsymbol{X}^{(h)},\boldsymbol{y})$ を行い，これを使って距離 $\boldsymbol{d}(\boldsymbol{X}^{(h)},\boldsymbol{y})$ を計算する；この距離のベクトルから最近傍集合 $\mathcal{I}(\boldsymbol{X}^{(h)},\boldsymbol{y})$ を求め，それを使ってスコアベクトル $\boldsymbol{s}(\boldsymbol{X}^{(h)},\boldsymbol{y})$ を得る．R 個のスコアベクトルの平均を使って，スコアベクトルの期待値 $\bar{\boldsymbol{s}}$ を近似する．この手法を**手法 C** と呼ぶ．

以上の提案手法を表 1 にまとめる．

3.1　分散の与え方

式 (3) に示すように，本研究では各データの分布を正規分布と仮定している．正規分布にはパラメータ μ_{ij}，および σ_{ij}^2 がある．平均パラメータ μ_{ij} は観測値，もしくは補完推定値を用いることは前述した．分散パラメータ σ_{ij}^2 の値は，欠損データに対する分散には $\sigma_{ij}^2=\sigma^2$ を用いることとし，観測されたデータには $\sigma_{ij}^2=(\lambda\sigma)^2$ を用いることとする．ただし，λ は 1 以下の正の定数である．節 4 で示す実験では，手法 A には $\lambda=0.5$ を用い，手法 B および手法 C では $\lambda=1$ を用いた．σ^2 の値を求めるために，観測データに対して，欠損補完法によって得られる補完推定値 $\tilde{x}_i^{(j)}$ を使って，観測値 $\tilde{x}_i^{(j)}$ との二乗誤差の平均を計算する．その結果を σ^2 と定める．

4. 実験結果

提案手法の評価のために，データセット Colon[1] を用いた．このデータセットは 2 クラス識別問題で，正例を 22 個，負例を 40 個含み，結腸癌か否かを分類するタスクである．このデータセットには欠損が含まれていないが，提案手法の評価のために無作為に一定割合選んだデータに欠損を挿入する．遺伝子選択には t 検定における上位 100 個の遺伝子を選択した．モンテカルロ法の繰り返し数は $R=100$ とした．欠損補完アルゴリズムは，MEANimpute を用いた．leave-one-out 解析を行い，正答率を計算した．この操作を 100 回繰り返し，平均の正答率で評価することとした．

3 つの提案手法，手法 A，手法 B，手法 C と，節 2 で述べた従来手法 Conventional を比較した．その結果を図 1 に示す．手法 C は，どの欠損の割合においても従来法よりも高い識別性能を示した．手法 B は従来法よりは高精度に予測したが，手法 C に及ばなかった．このことより，データ行列 $\boldsymbol{X}$ は，途中のステップで点として扱うより，最後のステップまで確率変数として扱うべきであることを示唆している．手法 A は，高速に計算できるが，手法 B や C ほど良い性能を示さなかった．これより，$\mathcal{F}(\boldsymbol{X},\boldsymbol{y})$ を $\mathcal{F}(\bar{\boldsymbol{X}},\boldsymbol{y})$ に置き換えてしまうのはよい近似にはならないことを意味している．欠損の割合が小さいとき従来法と一致するのは，そもそも手法 A は欠損がない場合の結果が従来法と同じになるように設計されているからである．

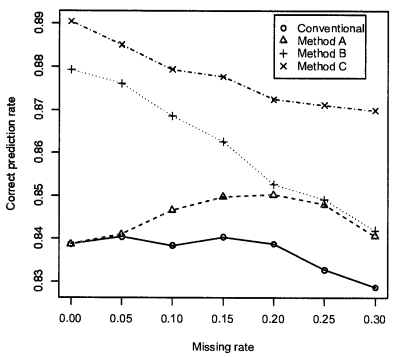

図 1　データセット Colon による 3 つの提案手法と従来手法の比較．手法 C が最も優れた性能を得ている．

表 1　各手法の概要

	従来法	手法 A	手法 B	手法 C
欠損補完	$\bar{\boldsymbol{X}}$ を計算.	$\bar{\boldsymbol{X}}$ を計算. $p(\boldsymbol{X})$ を決定.	$p(\boldsymbol{X})$ を決定. $\{\boldsymbol{X}^{(h)}\}_{h=1}^{R}$ を生成.	$p(\boldsymbol{X})$ を決定. $\{\boldsymbol{X}^{(h)}\}_{h=1}^{R}$ を生成.
遺伝子選択	$\mathcal{F}(\bar{\boldsymbol{X}}, \boldsymbol{y})$ を計算.	$\mathcal{F}(\bar{\boldsymbol{X}}, \boldsymbol{y})$ を計算.	各 h に対して $\mathcal{F}(\boldsymbol{X}^{(h)}, \boldsymbol{y})$ を計算.	各 h に対して $\mathcal{F}(\boldsymbol{X}^{(h)}, \boldsymbol{y})$ を計算.
距離計算	$\boldsymbol{d}(\bar{\boldsymbol{X}}, \boldsymbol{y})$ を計算.	式 (4) を使って距離を計算.	各 h に対して $\boldsymbol{d}(\boldsymbol{X}^{(h)}, \boldsymbol{y})$ を計算. 式 (6) を使って距離を計算.	各 h に対して $\boldsymbol{d}(\boldsymbol{X}^{(h)}, \boldsymbol{y})$ を計算.
投票	$\boldsymbol{s}(\bar{\boldsymbol{X}}, \boldsymbol{y})$ を計算.	式 (4) の値から スコアベクトルを計算.	式 (6) の値から スコアベクトルを計算.	各 h に対して $\boldsymbol{s}(\boldsymbol{X}^{(h)}, \boldsymbol{y})$ を計算. 式 (7) を使って スコアベクトルを計算.

このほかにも，3つのデータセットを使って実験を行い，欠損補完法も MEANimpute だけではなく，KNNimpute や SVDimpute を使った場合も試したが，いずれも同様な傾向が見られた．これらの結果については紙面の制約から割愛する．

5. おわりに

マイクロアレイデータの識別には，データを点としてではなく確率分布とみなして解析したほうが良好な性能が得られることを示した．本論文では，最近傍識別器を例にとってこの考え方を適用した算法の詳細を示したが，他の解析法にも同様なアプローチを考えることができる．また，提案法はマイクロアレイデータに限らず，雑音や欠損を含む他の問題にも有効かもしれない．著者らは，現在これらの課題に取り組んでいる．

参考文献

1) Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, Vol. 96, pp. 6745–6750 (1999).

2) Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D.: Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, Vol.95, pp.14863–14868 (1998).

3) Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*, Vol.286, pp.531–537 (1999).

4) Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S.: A Bayesian missing value estimation method for gene expression profile data, *Bioinformatics*, Vol. 19, pp. 2088–2096 (2003).

5) Toh, H. and Horimoto, K.: Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling, *Bioinformatics*, Vol.18, pp.287–297 (2002).

6) Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Bostein, D. and Altman, R.B.: Missing value estimation methods for DNA microarrays, *Bioinformatics*, Vol.17, pp.520–525 (2001).

7) Xiang, Q., Dai, X., Deng, Y., He, C., Wang, J., Feng, J. and Dai, Z.: Missing value imputation for microarray gene expression data using histone acetylation information, *BMC Bioinformatics*, Vol.9 (2008).