1M-7

遺伝子発現データ解析における遺伝子偏差を用いた前処理方法の提案

大上 雅史 † 越野 亮 †
† 石川工業高等専門学校 電子情報工学科

1 はじめに

近年，DNA マイクロアレイによって測定される遺伝子発現データを用いた，ガンの遺伝子診断が注目されている．例として，血液のガンとして知られる急性白血病を，急性骨髄性白血病(acute myeloid leukimia : AML)と急性リンパ性白血病(acute lymphoblastic leukimia : ALL)の 2 種類に分類するという問題がある．このような遺伝子発現データからのクラス分類問題はデータマイニングやパターン認識の問題に属し，Golub ら[1]を始めとしてこれまで数多くの研究が行われてきた．しかし，遺伝子発現データの属性数が膨大であることによる解析の難しさから，ゴールドスタンダードな解析手法が確立されていないのが現状である．

本研究では，遺伝子発現データの属性を絞り込むために遺伝子偏差を導入し，これを用いた前処理方法を提案する．さらに，急性白血病患者の遺伝子発現データにこの前処理を行った上で分類学習手法を適用することで，精度を向上させることを目的とする．

2 遺伝子発現データと分類学習

遺伝子の情報が細胞での構造や機能に変換される過程を発現という．この発現がどの程度行われているかを，DNA マイクロアレイを用いることで複数の遺伝子について一度に測定することができる．この測定によって得られた遺伝子発現データに対して分類学習手法を適用し，クラス分類を行う．ヒトの遺伝子数は約 22,000 個であるといわれているので，遺伝子発現データに含まれる遺伝子の数も数千～数万オーダになる．即ち遺伝子発現データは属性数が膨大な大規模問題に属する．

3 従来手法

遺伝子発現データのような属性数の大きいデータに対して分類学習手法を適用する場合，計算量が大きくなる上，冗長な属性や無関係な属性によって精度を落とす結果となるため，全体の属性から重要な属性を選択してデータを小さくする必要がある．属性選択の手法には，データの一般的特徴に基づいて属性を選び出すフィルタメソッドと，機械学習手法自身によって属性を選び出すラッパーメソッドがあるが，

Algorithm Relief

► n : データセットのサンプル数
► N : データセットの属性数
for $j := 1$ **to** N **do**
　$W_j := 0$
end for
for $i := 1$ **to** n **do**
　サンプル x_i を選択
　x_i の near hit x_h と near miss x_m を検索
　for $j := 1$ **to** N **do**
　　$W_j := W_j - \mathrm{diff}_j(x_i, x_h)/n + \mathrm{diff}_j(x_i, x_m)/n$
　end for
end for
for $j := 1$ **to** n **do**
　if $W_j >$ 閾値 **then** $S_0 := S_0 \cup \{$属性 $f_j\}$
end for
S_0 を出力

$$\mathrm{diff}_j(x_i, x_\ell) = \begin{cases} |x_{ij} - x_{\ell j}| & (f_j \text{が数値属性}) \\ 0 & x_{ij} = x_{\ell j} \ (f_j \text{が名義属性}) \\ 1 & x_{ij} \neq x_{\ell j} \ (f_j \text{が名義属性}) \end{cases}$$

near hit : 最も近い同クラスのサンプル
near miss : 最も近い異クラスのサンプル
（全属性を用いたユークリッド距離による）

図 1: Relief アルゴリズム[2]

ラッパーメソッドは，精度は良いが計算量が大きいので，大規模問題への適用は非現実的である．そこで，本研究では遺伝子発現データの属性選択に，フィルタメソッドを用いる．図 1 にフィルタメソッドの代表的手法である Relief [2] のアルゴリズムを示す．Relief は各属性に対して重み W_j を計算し，ユーザが与えた閾値を超える重みをもつ属性 f_j を選択するという手法である．なお，データはバイナリクラスに限定される．

4 提案手法

Relief はフィルタメソッドの中でも基本的な手法のひとつであるが，精度は良いとは言えず，また遺伝子発現データのような大規模問題に適用する場合には，ラッパーメソッドほどではないにしろ，計算量は大きくなってしまう．そこで，Relief よりも単純なアルゴリズムで計算量が小さい属性選択手法として，GDS（**G**ene **D**eviation **S**election）を提案する．

4.1 遺伝子偏差と GDS

以下の式で定義される σ_f を属性 f の遺伝子偏差と定義する．GDS では，学習データの全属性に対し各属性の遺伝子偏差を

Proposal of a pre-processing method using gene deviation on gene expression data analysis
Masahito OHUE †, Makoto KOSHINO †
† Ishikawa National College of Technology, Dept. of Electronics and Information Technology

求め，ユーザが与えた割合によって選択属性数を決定し，遺伝子偏差の大きい属性から順に属性を選択する．ただしデータはバイナリクラス $C \in \{A, B\}$ とする．

$$\sigma_f = \left| \frac{1}{n_A} \sum_{k_A=1}^{n_A} \left| x_{k_A f} - \left\langle x_f \right\rangle_A \right| - \frac{1}{n_B} \sum_{k_B=1}^{n_B} \left| x_{k_B f} - \left\langle x_f \right\rangle_B \right| \right|$$

ここで n_i はクラス $i \in C$ に属するサンプル数，f は各属性，k_i はクラス i の各サンプル，$x_{k_i f}$ はサンプル k_i・属性 f の値，$\left\langle x_f \right\rangle_i$ はクラス i のデータの平均を表す．遺伝子偏差という名前は，発現データの属性が各遺伝子であることに由来する．

4.2 GDS の性能評価

GDS の精度を調べるため，比較実験を行った．データセットは，UCI ML リポジトリの 2 種類のデータセット（pima, ionosphere）と，Broad Institute の 1 種類のデータセット（prostate）を使用し，評価は C4.5[3]を用いた決定木学習の 10 fold Cross Validation による分類精度とした．また，選択属性数は，データセットの属性数に応じて 1%～50%の中で決定した．なお，C4.5 の決定木学習には，C4.5 が予め実装されているデータマイニングツール Weka[4]を使用し，属性選択手法の Relief と GDS は，Microsoft Visual C++ 6.0 を用いて新たにプログラムを作成した．実験結果を表 1 に示す．

表 1 から，GDS は属性選択を行わない場合よりも精度が向上しており，Relief と同程度の精度が得られていることが分かる．特に，属性数の大きい prostate に対しては，Relief が良い効果を発揮できていないが，GDS では精度を向上させることに成功している．

表 1 : GDS の比較実験結果

	属性数	サンプル数	属性選択なし(%)	Relief (%)	GDS (%)	選択属性数
pima	8	768	72.8	74.1	74.7	4
ionosphere	34	351	90.9	92.6	91.5	10
prostate	12600	102	85.3	84.3	87.3	500

5 急性白血病クラス分類

Golub らの研究と同様に，急性白血病患者の遺伝子発現データを用いて AML/ALL のバイナリクラス分類実験を行った．実験方法，およびその結果を以下に示す．

5.1 実験

クラス分類実験に用いた遺伝子発現データは，属性数が 7109 であり，サンプル数は学習データが 38（AML : 11，ALL : 27），テストデータが 34（AML : 20，ALL : 14）である．この学習データに対し，Relief と GDS によって属性選択を行った 2 種の学習データを生成し，属性選択を行わない場合との比較を行った．このとき，選択属性数は全属性の約 1 %である 100 個とした．また精度は，C4.5 によって分類ルールを導き，得られた分類ルールをテストデータに適用したときの適合率とした．なお，実験に使用した計算機の性能は，CPU が Pentium 4 3.20GHz で，メモリは 512MB である．

5.2 結果

クラス分類の実験結果を表 2 に示す．表 2 から分かるように，Relief による属性選択では学習データの適合率は向上したものの，分類精度は 70.6%と，属性選択を行わない場合と同じであった．一方 GDS による属性選択では，分類精度が 97.1%と極めて良好な結果が得られた．また，計算時間についても，Relief による属性選択は 0.47 秒であったのに対し，GDS では 0.22 秒と，約 50%の計算時間削減に成功している．分類ルール導出までの一連の処理時間を見ても，属性選択を行わない場合の 1.69 秒に対し GDS では 0.24 秒であることから，GDS による属性選択は急性白血病クラス分類に対して，非常に効果的であることが分かる．

表 2 : 遺伝子発現データによる急性白血病クラス分類

	属性選択なし	Relief	GDS
学習データ(%)	97.4	100	100
テストデータ(%)	70.6	70.6	97.1
計算時間(sec)（属性選択 / 分類）	ー / 1.687	0.466 / 0.021	0.215 / 0.021

6 まとめ

遺伝子発現データからのクラス分類問題に対し，属性選択手法として GDS を提案し，実際に急性白血病患者の分類に GDS を用いることで，従来手法より計算時間を削減し，かつ精度を向上させることに成功した．今後の課題として，GDS をマルチクラスのデータに適用できるように拡張し，Relief をマルチクラスに拡張した Relief-F[5]との比較実験を行い，肺がんなどのマルチクラス分類問題に対し，遺伝子発現データによる分類を行うことが挙げられる．

References

[1] T.R.Golub and et al，“Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”，Science 286(15)，pp.531-537，1999．

[2] K.Kira and L.Rendell，“The feature selection problem: traditional methods and new algorithm”，Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92)，pp.129-134，1992．

[3] J.R.Quinlan，“C4.5: Programs for Machine Learning”，Morgan Kaufmann，1993．

[4] Weka（Wakikato Environment for Knouledge Analysis），http://www.cs.waikato.ac.nz/~ml/weka/

[5] I.Kononenko，“Estimating Attributes: Analysis and Extensions of RELIEF”，Proceedings European Con- ference on Machine Learning，1994．

[6] 三浦輝久，“Subset-Relief 法によるデータマイニングのための属性選択手法”，第 20 回人工知能学会全国大会，2A3-04，2006．