人工知能学会研究会資料
SGI-FPAI-B503-07

疾病有無判別のための相互情報量を用いたSNP組合せ探索手法

A method to identify disease-related SNP combinations using mutual information

真矢滋[1*]　小磯貴史[1†]　植野研[1‡]

[1]（株）東芝研究開発センターシステム技術ラボラトリー
[1] System Engineering Laboratory, Corporate Research & Development Center, Toshiba Corporation.

Abstract:
DNA sequences consist of about three billion base pairs. Although the genotypes of most bases are common within human being, SNPs are known to have different genotypes among individuals which may be related to a disease onset. In this paper, we consider the issue of classifying the disease presence or absence as accurately as possible based on the combinations of SNPs. Because the number of combinations of SNPs is huge, it may lead to difficulty to identify the proper combinations by the full search approach. In our proposed method, the basic idea is to select the SNPs which have the dependency with the disease using the mutual information considering the relationships among the SNP combinations efficiently. This process can reduce the candidates of combinations while maintaining the classification accuracy. We empirically demonstrate the effectiveness of our proposed method using real anonymised patient dataset (WTCCC-RA), showing that our methods performed better than comparative methods.

1　はじめに

人間の持つDNAはおよそ30億の塩基対から構成されているが，そのうち0.1%ほどが個人によりタイプが異なりＳＮＰ（一塩基多型）と呼ばれている．SNPと疾病の関連分析では，発症有と発症無の検体群のSNP情報を比較し，発症有無との関連が強いSNPを検出する．しかしながら，多くの疾病の発現には様々なSNPが関与しており，SNPを適切に検出することは困難である[6]．本稿の目的は疾病の有無をできるだけ正確に識別するためのSNPの条件に関する識別規則を検出することである．疾病を識別する規則を検出することで，疾病の発症に関するゲノムレベルでのメカニズムの解明につながる可能性がある．また識別規則と各検体のゲノム情報を比較することで将来的に疾病の発症予測に利用することも考えられる．これまでカイ二乗検定やフィッシャーの正確性検定を用いて単一のSNPを用いて疾病の有無を識別する解析が行われてきた．しかしながら単一のSNPに基づく場合には疾病との強い関連がある主効果がある場合には識別が可能であるが，複数のSNP間の相互作用を加味する必要があるような疾病には対応できない問題がある．

そこで複数のSNPの組合せによって疾病とＳＮＰの関連を分析する手法が提案されてきた．MDRでは予め指定された個数のSNPによる相互作用を同時に考慮することができる[3]．しかしながらMDRはSNPの組合せに関して全探索と交差検証を行いより頑健な組合せを検出するため計算時間に問題がある．またBOOSTではMDRと同様にSNPの組合せに関して全探索を行うが，データの表現をブール型に指定することで計算時間を効率化している[8]．これらの手法は全探索を行うため，３つ以上のSNP間の相互作用を把握するには計算時間の観点から不向きである．

一方で多くのSNPから疾病の識別有無と関連が高いSNPを選択することは変数選択問題の一種である．mRMRは相互情報量を用いて被説明変数と依存度が高く，かつ互いに冗長になりづらい説明変数を効率的に検出する方法である[7]．またPengらはmRMRを用いて遺伝子の選択を行っている[1]．また他にも相互情報量も基にしたSNPの探索手法は提案されている[4][5]．

これら相互情報量を基にした手法は複数のＳＮＰから構成されるSNPの組合せをただ一組だけ出力する．そのため選択するSNPの個数を増やしすぎると，組合せに含まれる全てのSNPに関する識別規則を満たす検

*連絡先：（株）東芝研究開発センターシステム技術ラボラトリー
〒212-8582 神奈川県川崎市幸区小向東芝町1
E-mail: shigeru1.maya@toshiba.co.jp
†E-mail: takashi.koiso@toshiba.co.jp
‡E-mail: ken.ueno@toshiba.co.jp

体数自体が減少し識別性能が悪化する問題がある．そのため，本稿では SNP の組合せを複数導出し，それらを組み合わせることで全体として識別性能が向上する疾病有無に関する識別規則を検出することを目指す．

しかしながら，ＳＮＰの組合せの場合の数は膨大であり，また同様にそれらの組合せの場合の数も膨大である．そこで本稿では相互情報量を用いて識別規則に含まれる可能性がある SNP の探索範囲を限定し，効率的な識別規則の導出を行う．

提案手法は大きく 2 つのステップから構成される．第 1 ステップでは SNP の組合せ間の関係を考慮した上で，相互情報量に基づき疾病との関連が高い SNP の選択を行い，SNP の探索範囲の絞り込みを行う．第 2 ステップでは第 1 ステップで選ばれた SNP に関して識別誤差に基づき実際にそれぞれの SNP を識別規則に組み込むべきか否かを決定する．

本稿の主な貢献は次の点である．ＳＮＰ間の相互作用を考慮しながら SNP の組合せを複数個導出し，それらを組み合わせることで疾病の有無をできるだけ正確に識別できる枠組みを提案したことである．また疾病に関する補助情報を利用していないため，様々な疾病の分析に利用可能であると考えられる．

本稿の構成は以下の通りである．まず 2 章において本稿で扱うデータセットについて述べる．3 章では本稿で対象とする識別規則の定義を行う．4 章では相互情報量を用いた変数選択の既存手法を紹介する．5 章では提案手法について述べる．6 章では実際の関節リウマチのデータを用いて実験を行い提案手法の有効性を示す．最後に 7 章で結論を述べる．

2　データセットについて

本節では扱うデータセットについて紹介を行う．データセットに属する検体数を N，SNP 数を M とする．

まず疾病の有無に関するデータセットについて紹介を行う．ベクトル $\boldsymbol{y} \in \{0,1\}^N$ は各検体の疾病の有無を表したものである．ベクトル $\boldsymbol{y}$ の第 i 番目の要素の値を $\boldsymbol{y}_i$ で表し，第 i 番目の検体が疾病を有する場合には $\boldsymbol{y}_i = 1$ とし，疾病を有さない場合には $\boldsymbol{y}_i = 0$ とする．

次に SNP の情報に関するデータセットについて述べる．それぞれの SNP には 3 種類のタイプ（遺伝子型）が存在する．それぞれメジャーホモ接合体 (AA)，ヘテロ接合体 (AB)，マイナーホモ接合体 (BB) と呼ばれている．本稿では各検体，各 SNP のタイプを行列 $\boldsymbol{X} \in \{0,1\}^{N\times 3M}$ で表す．第 n 番目の検体に関して第 m 番目の SNP のタイプが AA ならば $\boldsymbol{X}_{n,3(m-1)+1} = 1$，AB ならば $\boldsymbol{X}_{n,3(m-1)+2} = 1$，BB ならば $\boldsymbol{X}_{n,3(m-1)+3} = 1$ とする．ただし $\boldsymbol{X}_{i,j}$ は行列 $\boldsymbol{X}$ の (i,j) 番目の要素である．図 1 に各検体，各 SNP のタイプを行列 $\boldsymbol{X}$ の例を示す．

	SNP-1:AA	SNP-1:AB	SNP-1:BB	...	SNP-9:BB
ID:1	0	1	0	...	0
ID:2	1	0	0	...	1
ID:3	0	0	1	...	0
ID:4	0	0	1	...	0

図 1: 各検体，各 SNP のタイプを表現した行列 $\boldsymbol{X}$ の例．行列 $\boldsymbol{X}$ の要素の値が 1 となる部分が各検体の対応する SNP のタイプである．

	SNP-1:AA	SNP-1:AB	SNP-1:BB	...	SNP-9:BB
疾病要因1	0	1		...	1
疾病要因2	1	0	0	...	1

図 2: 疾病規則を表す行列 $\boldsymbol{Z}$ の例．疾病要因数は $K = 2$ であり，識別規則は “（SNP1 が AB かつ SNP9 が BB）または（SNP1 が AA かつ SNP9 が BB）ならば該当する検体は陽性”である．

3　疾病有無の識別規則

本稿で扱う疾病有無の識別規則を定義する．

まず疾病有無の識別規則として陽性の検体が同時に満たすべき SNP のタイプに関する規則を考える．例えば “（SNP1 が AA かつ SNP2 が AB かつ SNP3 が AB）ならば該当する検体は陽性である”という規則が挙げられる．この陽性の検体に特有の SNP のタイプに関する規則を “疾病要因 ”と呼ぶ．

また要因数を単一から複数にすることで識別精度が向上することが考えられる．本稿では疾病要因を複数 (K 個) 設定し，いずれかの疾病要因の条件を満たす場合に疾病ありと識別することを考える．疾病要因数 $K = 2$ の場合の例としては “（SNP1 が AA かつ SNP2 が AB かつ SNP3 が AB）または（SNP3 が AA かつ SNP4 が AB）ならば該当する検体は陽性である”といった規則が挙げられる．そして，いずれかの疾病要因を満たす場合に陽性と判定する．

次に複数の疾病要因における識別規則を行列（$\boldsymbol{Z} \in \{0,1\}^{K\times 3M}$) を用いて表現する．行列 $\boldsymbol{Z}$ の第 k 行が，k 番目の疾病要因の条件に含まれる SNP のタイプを示す．そのため，あり得る識別規則の総数は $\boldsymbol{Z}$ の各要素が $\{0,1\}$ を取る場合の数だけ存在する．図 2 に識別規則を表す行列 $\boldsymbol{Z}$ の例を示す．またアルゴリズム 1 に SNP 情報を示す X と識別規則を示す $\boldsymbol{Z}$ が与えられた場合に n 番目の検体の疾病の有無を識別する方法を示す．

Algorithm 1 検体の疾病有無の識別方法.

入力 : SNP 情報に関する行列 $\boldsymbol{X}$，疾病有無に関するベクトル $\boldsymbol{Y}$，疾病要因数 K，対象の検体 ID: n.
出力 : i 番目の検体の識別結果.
変数 L を 0 に初期化.
for $k = 1 \rightarrow K$ **do**
  if k 番目の疾病要因の条件を n 番目の検体の各 SNP のタイプが満たす. **then**
    $L \leftarrow L + 1$
  end if
end for
if $Z \geq 1$ **then**
  疾病ありと出力.
else
  疾病なしと出力.
end if

4 mRMR

本節では，SNP 間の相互作用を考慮した変数選択手法（*minimal-redundancy-maximal-relevance*: mRMR）の紹介 [7] を行う．なお，提案手法は mRMR を基に考案した．本節の目的は疾病の有無を示すベクトル $\boldsymbol{y}$ と SNP 情報を示す行列 $\boldsymbol{X}$ が与えられた場合に，SNP 間の相互作用を考慮しながら疾病と関連が強い SNP のタイプを指定した個数 (V 個) だけ選択することである．疾病と関連が高い SNP のタイプを選択する必要があるが，疾病の有無と依存関係が高い SNP のタイプのみを選択すると，冗長なＳＮＰのタイプを選択してしまう可能性がある．冗長な SNP のタイプとは既に選ばれた SNP のタイプと依存関係が極めて高く全体としての性能の向上に貢献していないものである．そのため疾病との関連が強く，かつ選ばれた SNP のタイプ間の冗長度が低い SNP を選択する必要がある．mRMR では上記の 2 つの点を考慮しながら変数選択を行う．第 m 番目の SNP のタイプに対応する確率変数を x_m とする．また確率変数 X と Y の相互情報量を $I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$ と定義する．ただし $p(x,y)$ は (x,y) の同時確率である．mRMR では逐次的に特徴の選択を行うため，現在 $v-1$ 個 SNP のタイプを既に選択し，v 個目を選択する場合を考える．現在までに選ばれた $v-1$ 個の SNP のタイプの集合を $\mathcal{S}_{v-1}$ とし，全 SNP のタイプの集合を $\mathcal{A}$ とする．このとき v 番目に選択する SNP のタイプ (x_v) は

$$\underset{x_v \in \mathcal{A} - \mathcal{S}_{v-1}}{\operatorname{argmax}} \left[I(x_j;Y) - \frac{1}{v-1} \sum_{x_i \in \mathcal{S}_{v-1}} I(x_v;x_i) \right] \quad (1)$$

とする．ただし Y は疾病の有無に関する確率変数を示す．式 (1) の第 1 項は SNP と疾病の有無の依存関係を表したものであり，第 2 項は既に選択された SNP との冗長度を表したものである．

5 提案手法

本節では検体に関する疾病の有無の情報と各 SNP のタイプが与えられた場合に，疾病の有無をできるだけ正確に判別する識別規則を検出する問題を扱う．識別規則は SNP のタイプの組合せの総数だけ存在するため数が膨大である．そこで提案手法では適切に SNP のタイプの探索範囲を限定することで識別精度と計算時間を両立することを目指す．提案手法は 2 つのステップから構成される．第 1 ステップでは相互情報量に基づく変数選択手法を用い，疾病と関連が高いと考えられる SNP のタイプを選抜する．第 2 ステップでは第 1 ステップで選抜された各 SNP のタイプに関して，識別誤差を基に識別規則に取り組むべきかの判定を行う．

5.1 第 1 ステップ

第 1 ステップでは識別規則を表す行列 $\boldsymbol{Z}$ の $3KM$ 個の要素から疾病との関連が高い V 個を効率的に選抜する．mRMR を用いることで疾病との関連が強くかつ冗長度が低い SNP のタイプを選抜できるが，mRMR では単一の組合せを想定しており疾病要因間の関係は考慮されていない．そこで疾病要因間の関係を考慮した変数選択法を提案する．

まず疾病要因間の関係を考慮した場合の第 k 番目の疾病要因に関する第 m 番目の SNP のタイプと疾病との関連度の導出方法を述べる．本稿では K 個の疾病要因を仮定しているが、k 番目の要因を除く $K-1$ 個の要因によって既に疾病ありと識別されている患者は k 番目の要因の内容に関わらず疾病ありと識別されてしまう．逆に言えば $K-1$ 個の要因で疾病なしと識別されている検体に関して m 番目の SNP のタイプと疾病の有無の依存関係を測るべきである．現時点での第 k 番目の要因を除く $K-1$ 個の要因で疾病なしと識別されている検体の ID の集合を $\boldsymbol{T}_k$ とする．そして第 k 番目の要因に関する第 m 番目の SNP のタイプと疾病との依存度を $I(\boldsymbol{Y}_{\boldsymbol{T}_k}, \boldsymbol{X}_{\boldsymbol{T}_k,m})$ とする．ただし，$I(\boldsymbol{Y}_{\boldsymbol{T}_k}, \boldsymbol{X}_{\boldsymbol{T}_k,m})$ は ID が $\boldsymbol{T}_k$ に属する検体に関して m 番目の SNP のタイプと疾病の有無に関する相互情報量である．

次に疾病要因間の関係を考慮した場合の第 k 番目の疾病要因に関する第 m 番目の SNP のタイプと第 i 番目の疾病要因に関する第 j 番目の SNP のタイプの冗長度を測る．一般に k 番目の要因を除く $K-1$ 個の要因で疾病なしと識別された検体と i 番目の要因を除く $K-1$ 個の要因で疾病なしと識別された検体は異なる．そこで両者の共通の検体に基づいて相互情報量を求め

表 1: 分割表の例

	陽性と診断	陰性と診断	合計
実際は陽性	a	b	$a+b$
実際は陰性	c	d	$c+d$
合計	$a+c$	$b+d$	$a+b+c+d$

る．それぞれの要因を除く $K-1$ 個の要因で疾病ありと識別されていない検体の集合を $\boldsymbol{T}_k$ と $\boldsymbol{T}_i$ とし，両者の集合の積集合を $\boldsymbol{T}_{k,i}$ とする．このとき，2 つの SNP のタイプ間の冗長度を

$$\frac{|\boldsymbol{T}_k \cap \boldsymbol{T}_i|}{|\boldsymbol{T}_k \cup \boldsymbol{T}_i|} I(\boldsymbol{X}_{T_{k,i},m}, \boldsymbol{X}_{T_{k,i},j}) \tag{2}$$

とする．ただし $I(\boldsymbol{X}_{T_{k,i},m}, \boldsymbol{X}_{T_{k,i},j})$ は ID が $\boldsymbol{T}_{k,i}$ に属する検体に関する m 番目の SNP のタイプと j 番目の SNP のタイプ間の相互情報量である．

以上から mRMR にならい v 番目に選抜すべき SNP のタイプ x_v は

$$\underset{x_v \in \mathcal{A}-\mathcal{S}_{v-1}}{\operatorname{argmax}} \Big[I(\boldsymbol{Y}_{\boldsymbol{T}_k}, \boldsymbol{X}_{\boldsymbol{T}_k,m}) - \frac{1}{v-1} \sum_{x_i \in \mathcal{S}_{v-1}} \frac{|\boldsymbol{T}_k \cap \boldsymbol{T}_i|}{|\boldsymbol{T}_k \cup \boldsymbol{T}_i|} I(\boldsymbol{X}_{T_{k,i},m}, \boldsymbol{X}_{T_{k,i},j}) \Big] \tag{3}$$

となる．式 (3) では v 番目の SNP のタイプを選択のため $(M-v+1)$ 個それぞれの SNP のタイプに対し，これまでに選択された $v-1$ 個の SNP のタイプとの相互情報量を計算する必要がある．また相互情報量の計算では最大 N 検体の情報を用いる．以上から第 1 ステップの計算量は $\mathcal{O}(\sum_{v=1}^{V} N(M-v+1)(v-1)) = \mathcal{O}(NMV^2)$ となる．

5.2 第 2 ステップ

本ステップでは第 1 ステップで選択された各疾病要因の SNP のタイプを実際に識別規則に組み込むかを判定する．第 1 ステップで選択された V 個の要素が行列 $\boldsymbol{Z}$ において 0 もしくは 1 の値を取る場合の数は 2^V 通りである．本ステップではこれら 2^V の通りの中で，識別誤差が最小となる組合せに $\boldsymbol{Z}$ を更新する．ただし第 2 ステップでは第 1 ステップで選択された V 個以外の SNP のタイプに関しては $\boldsymbol{Z}$ の要素の値を更新しない．

次に識別誤差の定義について述べる．行列 $\boldsymbol{Z}$ に対応して各検体はアルゴリズム 1 に基づいて疾病の有無が識別される．そして各検体が実際に陽性か否かに関して分割表 (図 1) が得られる．

図 1 では c，および b が誤識別した検体の数に相当する．実際の陽性の検体数と陰性の検体数に偏りがある場合も想定し識別誤差を分割表の値を用いて

$$\gamma c + (1-\gamma) b \tag{4}$$

Algorithm 2 識別規則の検出に関する提案手法.

1: 入力：SNP 情報に関する行列 $\boldsymbol{X}$，疾病の有無に関するベクトル $\boldsymbol{Y}$，パラメータ U, V.
2: 出力：識別規則を示す行列 $\boldsymbol{Z}$.
3: 識別規則を表す $\boldsymbol{Z}$ を初期化.
4: **for** $u = 1 \to U$ **do**
5: 　第 1 ステップ.
6: 　**for** $v = 1 \to V$ **do**
7: 　　式 (3) に基づき v 番目の要素 (x_v) を選択.
8: 　**end for**
9: 　第 2 ステップ.
10: 　2^V の要素の組合せから識別誤差 (式 (4)) が最小となるものに $\boldsymbol{Z}$ を更新.
11: **end for**

と定義する．ただし γ はパラメータである．第 2 ステップの計算時間は V 個の組合せに関して全探索を行うため計算量は $\mathcal{O}(2^V)$ となる．

提案手法では第 1 ステップと第 2 ステップを指定した回数 (U) だけ繰り返す．そして最終的な行列 $\boldsymbol{Z}$ を求め，疾病の有無の識別に活用する．アルゴリズム 2 に提案手法の全体の流れを示す．また計算量は $\mathcal{O}(U(NMV^2 + 2^V))$ となる．実験的には 6 章で述べる通り，V は KM に比べて非常に小さな値に設定しても良い性能ため，計算時間は行列 $\boldsymbol{X}$ のサイズに依存する．

6 実験

本節では WTCCC より提供を受けた関節リウマチの患者データ (WTCCC-RA) を分析し，提案手法の有効性を評価する．

6.1 データセット

本稿では WTCCC[9] より提供を受けた関節リウマチのデータセットを用いる．検体数は 3499 であり，1999 検体が疾病あり，1500 検体が疾病なしである．今回は免疫系染色体番号が 6 である SNP に注目をした．染色体番号が 6 である SNP は合計 31470 個存在するが，本稿では各 SNP 毎にフィッシャーの正確性検定を行い，p 値が低いものから 1000 個の SNP を選抜し実験に用いた．そのため本稿で用いたデータセットのサイズとしては検体数 (N) が 3499，SNP 数 (M) が 1000 である．なお得られたデータに欠損値はなく，全ての SNP に関して対応するタイプが取得でき，全ての検体の疾病の有無に関する情報は得られているとする．ただし，疾病なしの検体群に関しては各 SNP でマイナーホモの検体が存在しない場合は全てメジャーホモとして扱った．

表 2: 提案手法の実験結果 (V=2，V=4)

適合率	再現率	F 値
1.000	0.627	0.771

6.2 評価指標

評価指標として適合率，再現率，F 値を用いる．分割表を用いた場合の定義は 適合率 := $a/(a+c)$，再現率 := $a/(a+b)$，F 値 := $(2 \times$ 適合率 $\times$ 再現率$)/($適合率 + 再現率$)$ である [2]．これらの指標は疾病の有無を正しく識別できるほど大きな値を示す．

6.3 比較手法

本節では比較手法の紹介を行う．naive-one は $3M$ 個の各 SNP のタイプから全探索により識別誤差が最小となる 1 つの SNP のタイプを選択する手法である．naive-two-and は $3M$ 個の各 SNP のタイプから全探索により 2 つの SNP を選択し，2 つの SNP のタイプを同時に満たす場合に陽性であるとする識別規則に基いた場合の識別誤差が最小とするものを選択する手法である．naive-two-or は $3M$ 個の各 SNP のタイプから全探索により 2 つの SNP を選択し，2 つの SNP のタイプのうちいずれかを満たす場合に陽性であるとする識別規則に基づいた場合の識別誤差を最小とするものを選択する手法である．mRMR は提案手法の第 1 ステップに mRMR を用いたものである．第 2 ステップに関しては提案手法と同一である．

6.4 実験結果および考察

本節では提案手法および比較手法の実験結果を示す．なお識別誤差におけるパラメータは $\gamma = 0.43$ とした．また提案手法および mRMR において U は 8 に設定をした．表 2，表 3 に V を変化させた場合の提案手法の適合率，再現率，F 値を示す．ただし，$V = 2, 4$ の場合には疾病要因数を 1 から 6 まで変化させても評価指標の値は変化しなかった．また各疾病要因数で実際に使用した SNP タイプ数も併記する．ただし同じ SNP のタイプを異なる要因で用いた場合には 1 回として計測を行う．使用 SNP タイプ数は $V = 2, 4, 8$ の場合にそれぞれ $[1, 2, 2, 3, 3, 3]$，$[1, 3, 3, 3, 4, 4]$，$[1, 4, 5, 5, 7, 7]$ と増加した．

次に比較手法の結果を表 4 に示す．なお mRMR($V = 8$) においては疾病要因数を変化させた場合でも分割表の値は変わらず，従って各評価指標の値も同一であった．

表 3: 提案手法の実験結果 (V=8)

疾病要因数	適合率	再現率	F 値
1	1.000	0.627	0.771
2	1.000	0.627	0.771
3	1.000	0.651	0.789
4	1.000	0.926	0.962
5	1.000	0.956	0.978
6	1.000	0.961	0.981

表 4: 比較手法の実験結果

比較手法	適合率	再現率	F 値
naive-one	1.000	0.627	0.771
naive-two-and	1.000	0.627	0.771
naive-two-or	1.000	0.926	0.962
mRMR	1.000	0.627	0.771

比較手法について考察を行う．表 4 に示す通り 4 通りの比較手法の中では naive-two-or の性能が最も優れていた．naive-one と naive-two-and は疾病要因数を 1 に設定した場合とみなすことができる．同様に，naive-two-or は疾病要因数を 2 に設定したものである．naive-one，naive-two-and，naive-two-or では naive-two-or の結果が最もよい．これは疾病要因数を増加させることで識別精度が向上することを示唆している．一方で，疾病要因数を増やすと取り得る SNP の場合の数が膨大となり，探索が困難になる問題がある．そこで mRMR では探索範囲を限定し，識別規則を検出している．しかしながら，mRMR では第 1 ステップで SNP 要素を選抜する際に，SNP タイプ間の相互作用しか考慮せず，疾病要因間の相互作用を考慮できていないため性能が改善しなかったと考えられる．

次に提案手法の結果について考察を行う．提案手法では $V = 2, 4$ の場合には疾病要因数を増加させても評価指標の値が改善しておらず，また疾病要因数を増加させても使用 SNP タイプ数はほとんど変わらず疾病要因数を増やした場合の利点が生かせていない．この原因としては $V = 2, 4$ と少ないため，第 1 ステップで疾病要因間の関係を捉えるのに十分な SNP 要素を選抜できなかったと考えられる．一方で $V = 8$ の場合には 表 3 に示す通り疾病要因数を増加させるにつれ再現率と F 値の指標が改善していることが分かる．また使用ＳＮＰタイプ数も増加していくことが分かる．

次に複数の疾病要因を設定した利点について述べる．1999 の陽性検体のうち提案手法で 6 個の疾病要因を仮定した場合には 1923 検体を陽性と正しく識別できたが，6 個の疾病要因それぞれによって正しく陽性の検体した件数を表 5 に示す．ただし複数の疾病要因で陽性と識別された場合には，陽性と識別した疾病要因数の

表 5: 各疾病要因が陽性の検体を識別した件数

1 つ目の要因	2 つ目の要因	3 つ目の要因
326	472	26
4 つ目の要因	5 つ目の要因	6 つ目の要因
268	212	620

逆数だけ加算し，最終的に四捨五入を行った．表 5 から，6 つ目の疾病要因の値が最大だが他の疾病要因の値も比較的大きい．つまり 6 個の疾病要因のうち不要なものはなく，それぞれの疾病要因は疾病の識別に貢献できていると考えられる．提案手法では複数の疾病要因を設定し，第 1 ステップにおける SNP 選抜においても SNP 間の相互作用と共に疾病要因間の相互作用を両方考慮することができ，識別精度の向上した．

なお比較手法も含め全ての場合で適合率が 1.000 となったが，本稿で用いたデータは疾病なしの検体群に関して一部 2 種類の接合体タイプしか存在しなかっためと考えられる．各 SNP に関する接合体の判定方法 (遺伝子型判定技術) は発展段階であり，本稿で用いたデータセットが完全に人間の遺伝情報を表現しているかについては議論の余地がある．

7 結論

本論文では，複数の疾病要因を仮定した場合に疾病要因間の関係を考慮しつつ疾病の有無をできるだけ正確に識別する規則の効率的な検出方法について扱った．提案手法では，第 1 ステップで相互情報量を用いて疾病との関連が高い SNP のタイプを選抜し，第 2 ステップで選抜された SNP のタイプのうち実際に識別規則に組み込むかを決定する 2 段階から成る手法を提案した．第 1 ステップにおける SNP タイプの選抜では，通常の mRMR では SNP 間の相互関係しか扱えなかったが，提案手法では SNP 間の相互関係に加え疾病要因間の関係も考慮できる．その結果，実データを用いた実験において疾病要因数を増加させた場合に提案手法は比較手法より良い性能を示した．今後の課題としては他の疾患への適用や，疾病要因数や第 1 ステップで選択するＳＮＰのタイプの個数の決定方法が挙げられる．

謝辞

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113, 085475 and 090355.

参考文献

[1] C. Ding and H. Peng, “Minimum redundancy feature selection from microarray gene expression data,” *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[2] B. C. M. Fung, K. Wang, and M. Ester, “Hierarchical document clustering using frequent itemsets,” In *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM)*, vol. 3, pp. 59–70, 2003.

[3] L. W. Hahn, M. D. Ritchie, and J. H. Moore, “Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions,” *Bioinformatics*, vol. 19, no. 3, pp. 376–382, 2003.

[4] T. Hu, Y. Chen, J. W. Kiralis, R. L. Collins, C. Wejse, G. Sirugo, S. M. Williams, and J. H. Moore, “An information-gain approach to detecting three-way epistatic interactions in genetic association studies,” *Journal of the American Medical Informatics Association*, vol. 20, no. 4, pp. 630–636, 2013.

[5] S. Leem, H. H. Jeong, J. Lee, K. Wee, and K.-A. Sohn, “Fast detection of high-order epistatic interactions in genome-wide association studies using information theoretic measure,” *Computational Biology and Chemistry*, vol. 50, pp. 19-28, 2014.

[6] J. H. Moore and S. M. Williams, “Epistasis and its implications for personal genetics,” *American Journal of Human Genetics*,vol. 85, no. 3, pp. 309–320, 2009.

[7] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[8] X. Wan, C. Yang, Q. Yang, H. Xue, X. Fan, N. Tang, and W. Yu, “BOOST a fast approach to detecting gene-gene interactions in genome-wide case-control studies,” *American Journal of Human Genetics*,vol. 87, no. 3, pp. 325–340, 2010.

[9] The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature, 447(7145), 661–678. http://doi.org/10.1038/nature05911.