(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6253644号
(P6253644)

(45)発行日　平成29年12月27日(2017.12.27)　(24)登録日　平成29年12月8日(2017.12.8)

(51)Int.Cl.		F I		
G06F 17/30	(2006.01)	G06F	17/30	210D
G06F 19/24	(2011.01)	G06F	17/30	170F
G06N 99/00	(2010.01)	G06F	19/24	
		G06N	99/00	153

請求項の数 15　(全 22 頁)

(21)出願番号　特願2015-517783(P2015-517783)
(86)(22)出願日　平成25年6月21日(2013.6.21)
(65)公表番号　特表2015-525413(P2015-525413A)
(43)公表日　平成27年9月3日(2015.9.3)
(86)国際出願番号　PCT/EP2013/062980
(87)国際公開番号　WO2013/190084
(87)国際公開日　平成25年12月27日(2013.12.27)
審査請求日　平成28年6月20日(2016.6.20)
(31)優先権主張番号　61/662,792
(32)優先日　平成24年6月21日(2012.6.21)
(33)優先権主張国　米国(US)

(73)特許権者　500586875
フィリップ　モリス　プロダクツ　エス　アー
スイス国　２０００　ヌーシャテル　ケ　ジャンルノー　３
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100113413
弁理士　森下　夏樹
(74)代理人　100181674
弁理士　飯田　貴敏
(74)代理人　100181641
弁理士　石川　大輔
(74)代理人　230113332
弁護士　山本　健策

最終頁に続く

(54)【発明の名称】統合バイアス補正およびクラス予測を用いてバイオマーカシグネチャを生成するためのシステムおよび方法

(57)【特許請求の範囲】
【請求項１】
プロセッサによって実行される、２つ以上のクラスにデータセットを分類するコンピュータ実装方法であって、前記方法は、
（ａ）トレーニングデータセットおよびトレーニングクラスセットを受信することであって、前記トレーニングクラスセットは、既知のラベルのセットを含み、各既知のラベルは、前記トレーニングデータセット中の各要素と関連付けられるクラスを識別する、ことと、
（ｂ）テストデータセットを受信することと、
（ｃ）前記トレーニングデータセットおよび前記トレーニングクラスセットに第１の機械学習技法を適用することによって、前記トレーニングデータセットについての第１の分類器を生成することと、
（ｄ）前記第１の分類器に従って、前記テストデータセット中の要素を分類することによって、第１のテストクラスセットを生成することと、
（ｅ）トレーニングクラス重心のセットの中心に対応する量だけ前記トレーニングデータセット中の前記要素を偏移させることによって、前記トレーニングデータセットを変換することであって、各トレーニングクラス重心は、前記トレーニングデータセット中の要素のサブセットの中心を表す、ことと、
（ｆ）複数の反復の各々について、
（ｉ）テストクラス重心のセットの中心に対応する量だけ前記テストデータセット中

の前記要素を偏移させることによって、前記テストデータセットを変換することであって、各テストクラス重心は、前記テストデータセット中の要素のサブセットの中心を表す、こと、

（ｉｉ）第２の分類器に従って前記変換されたテストデータセット中の前記要素を分類することによって、第２のテストクラスセットを生成することであって、前記第２の分類器は、前記変換されたトレーニングデータセットおよび前記トレーニングクラスセットに第２の機械学習技法を適用することによって生成される、こと、および、

（ｉｉｉ）前記第１のテストクラスセットと前記第２のテストクラスセットとが異なる場合、前記第２のテストクラスセットを前記第１のテストクラスセットとして記憶し、前記変換されたテストデータセットを前記テストデータセットとして記憶し、ステップ（ｉ）に戻ること

を行うこと

と

を含む、方法。

【請求項２】

前記第１のテストクラスセットと前記第２のテストクラスセットとが異ならない場合に、前記第２のテストクラスセットを出力することをさらに含む、請求項１に記載の方法。

【請求項３】

前記トレーニングデータセットの前記要素は、疾患を有する患者についての、前記疾患に耐性がある患者についての、または、前記疾患がない患者についての遺伝子発現データを表す、請求項１～２のいずれかに記載の方法。

【請求項４】

前記トレーニングデータセットは、集約データセット中のサンプルのランダムなサブセットから形成され、前記テストデータセットは、前記集約データセット中のサンプルの残っているサブセットから形成される、請求項１～３のいずれかに記載の方法。

【請求項５】

ステップ（ｅ）における前記偏移させることは、前記変換されたトレーニングデータセットを取得するように、前記トレーニングデータセットに回転、シアー、線形変換、または、非線形変換を適用することを含む、請求項１～４のいずれかに記載の方法。

【請求項６】

ステップ（ｉ）における前記偏移させることは、前記変換されたテストデータセットを取得するように、前記テストデータセットに回転、シアー、線形変換、または、非線形変換を適用することを含む、請求項１～５のいずれかに記載の方法。

【請求項７】

前記テストデータセットは、既知のラベルのテストセットを含み、各既知のラベルは、前記テストデータセット中の各要素と関連付けられるクラスを識別し、

前記第１のテストクラスセットは、前記テストデータセットについての予測されるラベルのセットを含み、

前記第２のテストクラスセットは、前記変換されたテストデータセットについての予測されるラベルのセットを含む、

請求項１～６のいずれかに記載の方法。

【請求項８】

前記複数の反復の各々について、前記第１のテストクラスセットを前記第２のテストクラスセットと比較することをさらに含む、請求項１～７のいずれかに記載の方法。

【請求項９】

前記第１の機械学習技法と前記第２の機械学習技法は同一である、請求項１～８のいずれかに記載の方法。

【請求項１０】

ステップ（ｅ）での前記変換は、ステップ（ｉ）の同一の変換を適用することによって行われる、請求項１～９のいずれかに記載の方法。

【請求項１１】

前記第２のテストクラスセットを表示デバイス、印刷デバイス、または、記憶デバイスに提供することをさらに含む、請求項１～１０のいずれかに記載の方法。
【請求項１２】
前記第１のテストクラスセットおよび前記第２のテストクラスセットは、前記第１のテストクラスセットの任意の要素が前記第２のテストクラスセットの対応する要素と異なる場合に、異なる、請求項１～１１のいずれかに記載の方法。
【請求項１３】
前記第２のテストクラスセットは、前記変換されたテストデータセットについての予測されるラベルのセットを含み、前記方法は、予測されるラベルの総数によって除算された前記第２のテストクラスセット中の正確な予測されるラベルの数を表す性能測定基準を計算することによって、前記第２の分類器を評価することをさらに含む、請求項１～１２のいずれかに記載の方法。
【請求項１４】
コンピュータ可読命令を記憶したコンピュータ可読記憶媒体であって、前記コンピュータ可読命令は、少なくとも１つのプロセッサを備えるコンピュータ化システムにおいて実行される場合、前記少なくとも１つのプロセッサに請求項１～１３のいずれかに記載の方法を実行させる、コンピュータ可読記憶媒体。
【請求項１５】
非一時的なコンピュータ可読命令を伴って構成される少なくとも１つのプロセッサを備えるコンピュータ化システムであって、前記非一時的なコンピュータ可読命令は、実行される場合、前記プロセッサに請求項１～１３のいずれかに記載の方法を実行させる、コンピュータ化システム。
【発明の詳細な説明】
【背景技術】
【０００１】
関連出願への参照
本願は、米国仮特許出願第６１／６６２，７９２号（発明の名称「Ｓｙｓｔｅｍｓ ａｎｄ Ｍｅｔｈｏｄｓ ｆｏｒ Ｇｅｎｅｒａｔｉｎｇ Ｂｉｏｍａｒｋｅｒ Ｓｉｇｎａｔｕｒｅｓ ｗｉｔｈ Ｉｎｔｅｇｒａｔｅｄ Ｂｉａｓ Ｃｏｒｒｅｃｔｉｏｎ ａｎｄ Ｃｌａｓｓ Ｐｒｅｄｉｃｔｉｏｎ」、２０１２年６月２１日出願）に対する３５Ｕ．Ｓ．Ｃ § １１９の下での優先権を主張し、それは、本明細書にその全体が援用される。
【０００２】
背景
生物医学分野において、特定の生物学的状態を示す物質、すなわち、バイオマーカを識別することが重要である。ゲノミクスおよびプロテオミクスの新しい技術が出現するにつれて、バイオマーカは、生物学的発見、薬剤開発、および、ヘルスケアにおいてますます重要になりつつある。バイオマーカは、多くの疾患の診断および予後のためだけではなく、治療法の開発のための基礎を理解するためにも有用である。バイオマーカの成功した効果的な識別は、新薬開発プロセスを加速させることができる。診断および予後と治療法との組み合わせによって、バイオマーカ識別はまた、現在の薬物治療の品質を向上し、したがって、薬理遺伝学、薬理ゲノム学、および、薬理プロテオミクスの使用において重要な役割を果たす。
【０００３】
高スループットスクリーニングを含むゲノムおよびプロテオームの分析は、細胞において発現させられるタンパク質の数および形態に関する豊富な情報を供給し、各細胞について、特定の細胞状態の特性を示す発現させられたタンパク質のプロファイルを識別する潜在的可能性を提供する。特定の場合において、この細胞状態は、疾患と関連付けられる異常生理学的反応の特性を示し得る。結果として、疾患を有する患者からの細胞状態を識別し、それを正常な患者からの対応する細胞の細胞状態と比較することによって、疾患を診

断して治療する機会を提供することができる。

【０００４】

これらの高スループットスクリーニング技法は、遺伝子発現情報の大量のデータセットを提供する。研究者らは、個人の多様な集団について再現可能に診断するパターンにこれらのデータセットを組織化するための方法を開発しようとしてきた。１つのアプローチは、複合データセットを形成するように複数のソースからのデータをプールし、次いで、データセットを発見／トレーニングセットおよびテスト／検証セットに分割することであった。しかしながら、転写プロファイリングデータおよびタンパク質発現プロファイリングデータは両方とも、しばしば、利用可能な数のサンプルに対する多数の変数によって特徴付けられる。

【０００５】

患者または対照の群からの検体の発現プロファイルの間の観察された差異は、典型的に、疾患または対照の集団内の生物学的変動または未知のサブ表現型、研究プロトコルにおける差異による部位特異的なバイアス、検体の取り扱い、器具条件（例えば、チップバッチ等）における差異によるバイアス、および、測定誤差による変動を含むいくつかの要因によって、弱められる。いくつかの技法は、データサンプルにおけるバイアスを補正しようとする（例えば、別のクラスよりもむしろ、データセットにおいて表されるサンプルの１つのクラスを有することに起因し得る）。

【０００６】

いくつかのコンピュータベースの方法が、疾患および対照のサンプルの間の差異を最も良く説明する一組の特徴（マーカ）を見出すために開発されてきた。いくつかの初期の方法は、ＬＩＭＭＡ、乳癌に関するバイオマーカを識別するためのＦＤＡ承認マンマプリント技法、ロジスティック回帰技法、および、サポートベクトルマシン（ＳＶＭ）等の機械学習方法のような統計的テストを含んでいた。概して、機械学習の視点から、バイオマーカの選択は、典型的に、分類タスクについての特徴選択問題である。しかしながら、これらの初期の解決策は、いくつかの不利点に直面した。これらの技法によって生成されるシグネチャは、しばしば、対象の包含および除外が異なるシグネチャにつながり得るので、再現可能ではなかった。これらの初期の解決策はまた、多くの偽陽性シグネチャを生成し、小サンプルサイズおよび高次元を有するデータセットに作用するので、ロバストではなかった。

【０００７】

したがって、臨床的な診断および／または予後についてのバイオマーカを識別するため、より一般的には、データセットの中の要素を２つ以上のクラスに分類するために使用されることができるデータマーカを識別するための改良型技法の必要性がある。

【発明の概要】

【課題を解決するための手段】

【０００８】

出願人らは、既存のコンピュータベースの方法が、クラス予測技法とは別にバイアス補正技法を不利に適用することを認識している。本明細書で説明されるコンピュータシステムおよびコンピュータプログラム製品は、バイオマーカおよび他のデータ分類適用において改善された分類性能を達成し得る、バイアス補正およびクラス予測への統合アプローチを適用する方法を実装する。特定すると、本明細書で開示されるコンピュータ実装方法は、バイアス補正およびクラス予測への反復アプローチを採用する。コンピュータ実装方法の種々の実施形態において、システム中の少なくとも１つのプロセッサが、トレーニングデータセットおよびトレーニングクラスセットを受信し、そのトレーニングクラスセットは、トレーニングデータセットの中の要素の各々と関連付けられるクラスを識別する。システム中のプロセッサはまた、テストデータセットを受信する。プロセッサは、機械学習技法をトレーニングデータセットおよびトレーニングクラスセットに適用することによって、トレーニングデータセットについての第１の分類器を生成し、第１の分類器に従ってテストデータセット中の要素を分類することによって、第１のテストクラスセットを生成

する。複数の反復の各々について、プロセッサは、トレーニングクラスセットとテストクラスセットとのうちの少なくとも１つに基づいて、トレーニングデータセットを変換し、以前のステップの変換を適用することによって、テストデータセットを変換し、変換されたトレーニングデータセットおよびトレーニングクラスセットに機械学習技法を適用することによって、変換されたトレーニングデータセットについての第２の分類器を生成し、第２の分類器に従って、変換されたテストデータセット中の要素を分類することによって、第２のテストクラスセットを生成する。プロセッサはまた、第１のテストクラスセットと第２のテストクラスセットとを比較し、第１のテストクラスセットと第２のテストクラスセットとが異なる場合、プロセッサは、第２のクラスセットを第１のクラスセットとして記憶し、変換されたテストデータセットをテストデータセットとして記憶し、反復の開始に戻る。本発明のコンピュータシステムは、上記で説明されるような方法およびその種々の実施形態を実装するための手段を備える。

例えば、本発明は、下記の項目を提供する。

（項目１）

プロセッサによって実行される、２つ以上のクラスにデータセットを分類するコンピュータ実装方法であって、前記方法は、

（ａ）トレーニングデータセットおよびトレーニングクラスセットを受信するステップであって、前記トレーニングクラスセットは、前記トレーニングデータセット中の要素の各々と関連付けられるクラスを識別する、ステップと、

（ｂ）テストデータセットを受信するステップと、

（ｃ）前記トレーニングデータセットおよび前記トレーニングクラスセットに機械学習技法を適用することによって、前記トレーニングデータセットについての第１の分類器を生成するステップと、

（ｄ）前記第１の分類器に従って、前記テストデータセット中の要素を分類することによって、第１のテストクラスセットを生成するステップと、

（ｅ）複数の反復の各々について、

（ｉ）前記トレーニングクラスセットおよび前記テストクラスセットのうちの少なくとも１つに基づいて、前記トレーニングデータセットを変換するステップと、

（ｉｉ）前記テストデータセットを変換するステップと、

（ｉｉｉ）前記変換されたトレーニングデータセットおよび前記トレーニングクラスセットに基づいて、第２の分類器に従って前記変換されたテストデータセット中の前記要素を分類することによって、第２のテストクラスセットを生成するステップと、

（ｉｖ）前記第１のテストクラスセットと前記第２のテストクラスセットとが異なる場合、前記第２のクラスセットを前記第１のクラスセットとして記憶し、前記変換されたテストデータセットを前記テストデータセットとして記憶し、ステップ（ｉ）に戻るステップと

を含む、方法。

（項目２）

前記第１のテストクラスセットと前記第２のテストクラスセットとが異ならない場合に、前記第２のクラスセットを出力するステップをさらに含む、項目１に記載の方法。

（項目３）

前記トレーニングデータセットの要素は、疾患を有する患者についての、前記疾患に耐性がある患者についての、または、前記疾患がない患者についての遺伝子発現データを表す、項目１～２のいずれかに記載の方法。

（項目４）

前記トレーニングデータセットおよび前記テストデータセットは、集約データセット中のサンプルを前記トレーニングデータセットまたは前記テストデータセットにランダムに割り当てることによって、生成される、項目１～３のいずれかに記載の方法。

（項目５）

ステップ（ｉ）、ステップ（ｉｉ）、または、ステップ（ｉ）とステップ（ｉｉ）との両

方の前記変換は、前記データセットの重心に基づいて前記データセットの要素を調整することによって、バイアス補正技法を行うステップを含む、項目１～４のいずれかに記載の方法。
（項目６）
前記バイアス補正技法は、前記データセットの各要素から前記重心の成分を差し引くステップを含む、項目５に記載の方法。
（項目７）
ステップ（ｉ）、ステップ（ｉｉ）、または、ステップ（ｉ）とステップ（ｉｉ）との両方における前記変換は、回転、シアー、線形変換、または、非線形変換を適用するステップを含む、項目１～６のいずれかに記載の方法。
（項目８）
前記複数の反復の各々について、前記第１のテストクラスセットを前記第２のテストクラスセットと比較するステップをさらに含む、項目１～７のいずれかに記載の方法。
（項目９）
前記複数の反復の各々について、前記変換されたトレーニングデータセットおよび前記トレーニングデータセットに機械学習技法を適用することによって、前記変換されたトレーニングデータセットについての前記第２の分類器を生成するステップをさらに含む、項目１～８のいずれかに記載の方法。
（項目１０）
ステップ（ｉｉ）での前記変換は、ステップ（ｉ）の同一の変換を適用することによって行われる、項目１～９のいずれかに記載の方法。
（項目１１）
前記第２のテストクラスセットを表示デバイス、印刷デバイス、または、記憶デバイスに提供するステップをさらに含む、項目１～１０のいずれかに記載の方法。
（項目１２）
前記第１のテストクラスセットおよび前記第２のテストクラスセットは、前記第１のテストクラスセットの任意の要素が前記第２のテストクラスセットの対応する要素と異なる場合に、異なる、項目１～１１のいずれかに記載の方法。
（項目１３）
エラー率に基づいて、前記第２の分類器の性能測定基準を計算するステップをさらに含む、項目１～１２のいずれかに記載の方法。
（項目１４）
コンピュータ可読命令を備えるコンピュータプログラム製品であって、前記コンピュータ可読命令は、少なくとも１つのプロセッサを備えるコンピュータ化システムにおいて実行される場合、前記少なくとも１つのプロセッサに項目１～１３のいずれかに記載の方法の１つ以上のステップを実行させる、コンピュータプログラム製品。
（項目１５）
非一時的なコンピュータ可読命令を伴って構成される少なくとも１つのプロセッサを備えるコンピュータ化システムであって、前記非一時的なコンピュータ可読命令は、実行される場合、前記プロセッサに項目１～１３のいずれかに記載の方法を実行させる非一時的なコンピュータ可読命令を伴って構成される少なくとも１つのプロセッサを備える、コンピュータ化システム。
【０００９】
上記で説明される方法の特定の実施形態において、本方法はさらに、第１のテストクラスセットと第２のテストクラスセットとが異ならない場合に、第２のクラスセットを出力するステップを含む。特定すると、上記で説明されるような反復は、第１のテストクラスセットおよび第２のテストクラスセットが収束し、かつ、予測された分類の間に差異がなくなるまで、繰り返され得る。上記で説明される方法の特定の実施形態において、トレーニングデータセットの要素は、疾患を有する患者について、疾患に耐性がある患者について、または、疾患がない患者についての遺伝子発現データを表す。トレーニングクラスセ

ットの要素は、トレーニングデータセット中のデータサンプルについての既知のクラス識別子に対応し得る。例えば、クラス識別子は、「疾患陽性」、「疾患免疫性」、または、「疾患なし」等のカテゴリを含み得る。

【００１０】

上記で説明される方法の特定の実施形態において、トレーニングデータセットおよびテストデータセットは、集約データセット中のサンプルをトレーニングデータセットまたはテストデータセットにランダムに割り当てることによって、生成される。集約データセットをトレーニングデータセットとテストデータセットとにランダムに分割することが、クラスを予測してロバストな遺伝子シグネチャを生成するために望ましくあり得る。さらに、集約データセットのサンプルは、分割の前に破棄され得るか、または、トレーニングデータセットあるいはテストデータセットのサンプルは、分割後に破棄され得る。上記で説明される方法の特定の実施形態において、トレーニングデータセットを変換するステップ、テストデータセットを変換するステップ、または、トレーニングデータセットを変換するステップとテストデータセットを変換するステップとの両方は、データセットの重心に基づいてデータセットの要素を調整することによって、バイアス補正技法を行うステップを含む。変換は、トレーニングクラスセットに基づいて変換を定義し得る変換関数に従って行われる。上記で説明される方法の特定の実施形態において、バイアス補正技法は、データセットの各要素から重心の成分を差し引くステップを含む。例えば、バイアス補正技法の結果は、データセットにおいて表される各クラスの重心を考慮することによって、トレーニングデータセット、テストデータセット、または、トレーニングデータセットおよびテストデータセットの両方の各要素が、「再び中心に置かれる」ことであり得る。上記で説明される方法の特定の実施形態において、トレーニングデータセットを変換するステップ、テストデータセットを変換するステップ、または、トレーニングデータセットを変換するステップとテストデータセットを変換するステップとの両方は、回転、シアー（ｓｈｅａｒ）、線形変換、または、非線形変換を適用するステップを含む。

【００１１】

上記で説明される方法の特定の実施形態において、本方法はさらに、複数の反復の各々について、第１のテストクラスセットと第２のテストクラスセットとを比較するステップを含む。比較の結果として、第１のテストクラスセットおよび第２のテストクラスセットは、第１のテストクラスセットの任意の単一の要素が第２のテストクラスセットの対応する要素とは異なる場合に、異なると言われ得る。概して、第１のテストクラスセット中の少なくとも所定の数の要素が第２のテストクラスセット中の対応する要素と異なる場合に、第１のテストクラスセットと第２のテストクラスセットとが異なると言われ得るように、閾値が設定され得る。

【００１２】

上記で説明される方法の特定の実施形態において、本方法はさらに、複数の反復の各々について、変換されたトレーニングデータセットおよびトレーニングデータセットに機械学習技法を適用することによって、変換されたトレーニングデータセットについての第２の分類器を生成するステップを含む。上記で説明される方法の特定の実施形態において、テストデータセットの変換は、トレーニングデータセットを変換するステップの変換と同一の変換を伴う。上記で説明される方法の特定の実施形態において、本方法はさらに、表示デバイス、印刷デバイス、または、記憶デバイスに第２のテストクラスセットを提供するステップを含む。上記で説明される方法の特定の実施形態において、本方法はさらに、エラー率に基づいて、第２の分類器の性能測定基準を計算するステップを含む。特定の実施形態において、限定されないが、線形判別分析（ＬＤＡ）、ロジスティック回帰、サポートベクトルマシン、ナイーブベイズ分類器等の線形分類器が好ましい。

【００１３】

本発明のコンピュータシステムは、上記で説明されるような方法の種々の実施形態を実装するための手段を備える。例えば、コンピュータプログラム製品が説明され、本製品は、少なくとも１つのプロセッサを備えるコンピュータ化システムにおいて実行される場合

、上記で説明される方法のうちのいずれかの１つ以上のステップをプロセッサに実行させるコンピュータ可読命令を備える。別の例において、コンピュータ化システムが説明され、本システムは、実行される場合、上記で説明される方法のうちのいずれかをプロセッサに実行させる非一時的なコンピュータ可読命令を伴って構成されるプロセッサを備える。本明細書で説明されるコンピュータプログラム製品およびコンピュータ化方法は、１つ以上のプロセッサを各々が含む１つ以上のコンピューティングデバイスを有するコンピュータ化システムにおいて実装され得る。概して、本明細書で説明されるコンピュータ化システムは、本明細書で説明されるコンピュータ化方法のうちの１つ以上を実行するようにハードウェア、ファームウェア、および、ソフトウェアを伴って構成されるコンピュータ、マイクロプロセッサ、論理デバイス、または、他のデバイスもしくはプロセッサ等の、プロセッサまたはデバイスを含む１つ以上のエンジンを備え得る。これらのエンジンのうちのいずれか１つ以上は、いずれか１つ以上の他のエンジンから物理的に分離可能であり得るか、または、共通のまたは異なる回路基板上の別個のプロセッサ等の、複数の物理的に分離可能な構成要素を含み得る。本発明のコンピュータシステムは、上記で説明されるような方法およびその種々の実施形態を実装するための手段を備える。エンジンは、随時、相互接続され得、さらに、随時、摂動データベース、測定可能値データベース、実験データのデータベース、および、文献データベースを含む１つ以上のデータベースに接続され得る。本明細書で説明されるコンピュータ化システムは、ネットワークインターフェースを通して通信する１つ以上のプロセッサおよびエンジンを有する分散型コンピュータ化システムを含み得る。そのような実装は、複数の通信システムにわたる分散型計算のために適切であり得る。

【図面の簡単な説明】

【００１４】

本開示のさらなる特徴、その性質、および、種々の利点は、類似参照文字が全体を通して類似部分を指す添付図面と関連して検討される下記の詳細な説明を考慮すると明白になる。

【００１５】

【図１】図１は、１つ以上のバイオマーカシグネチャを識別するための例示的なシステムを描写する。

【図２】図２は、データセット中の要素の分類を図示する。

【図３－１】図３は、データセットを分類するための例示的なプロセスの流れ図である。

【図３－２】図３は、データセットを分類するための例示的なプロセスの流れ図である。

【図４】図４は、図１のシステムの構成要素のうちのいずれか等のコンピューティングデバイスのブロック図である。

【図５】図５は、トレーニングデータセット中の遺伝子シグネチャのヒートマップである。

【発明を実施するための形態】

【００１６】

本明細書で説明されるシステムおよび方法の全体的な理解を提供するために、ここで、遺伝子バイオマーカシグネチャを識別するためのシステムおよび方法を含む特定の例証的実施形態が、説明される。しかしながら、本明細書で説明されるシステム、コンピュータプログラム製品、および、方法は、任意のデータ分類適用等の他の好適な適用のために適合させられかつ修正され得、そのような他の追加および修正は、その範囲から逸脱しないことが、当業者によって理解される。概して、本明細書で説明されるコンピュータ化システムは、本明細書で説明されるコンピュータ化方法のうちの１つ以上を実行するようにハードウェア、ファームウェア、および、ソフトウェアを伴って構成されるコンピュータ、マイクロプロセッサ、または、論理デバイス等の１つ以上のエンジン、プロセッサ、または、デバイスを備え得る。

【００１７】

図１は、本明細書で開示される分類技法が実装され得る、１つ以上のバイオマーカシグ

ネチャを識別するための例示的なシステム１００を描写する。システム１００は、バイオマーカジェネレータ１０２と、バイオマーカコンソリデータ１０４とを含む。システム１００はさらに、バイオマーカジェネレータ１０２およびバイオマーカコンソリデータ１０４の動作の特定の局面を制御するための中央制御装置（ＣＣＵ）１０１を含む。動作中に、遺伝子発現データ等のデータが、バイオマーカジェネレータ１０２で受信される。バイオマーカジェネレータ１０２は、複数の候補バイオマーカおよび対応するエラー率を生成するようにデータを処理する。バイオマーカコンソリデータ１０４は、これらの候補バイオマーカおよびエラー率を受信し、最適な性能尺度およびサイズを有する好適なバイオマーカを選択する。

【００１８】

バイオマーカジェネレータ１０２は、データを処理して一組の候補バイオマーカおよび候補エラー率を生成するためのいくつかの構成要素を含む。特定すると、バイオマーカジェネレータは、データをトレーニングデータセットとテストデータセットとに分割するためのデータ前処理エンジン１１０を含む。バイオマーカジェネレータ１０２は、トレーニングデータセットおよびテストデータセットを受信してテストデータセットの要素を２つ以上のクラス（例えば、罹患および非罹患、感染しやすい、および、免疫がある、および、罹患等）のうちの１つに分類するための分類エンジン１１４を含む。バイオマーカジェネレータ１０２は、データ前処理エンジン１１０によって選択されるテストデータに適用される場合の分類器の性能を決定するための分類器性能監視エンジン１１６を含む。分類器性能監視エンジン１１６は、分類器（例えば、分類にとって最も重要であるデータセットの要素の成分）に基づいて候補バイオマーカを識別し、１つ以上の候補バイオマーカについて、候補エラー率を含み得る性能尺度を生成する。バイオマーカジェネレータ１０２はさらに、１つ以上の候補バイオマーカおよび候補性能尺度を記憶するためのバイオマーカ記憶部１１８を含む。

【００１９】

バイオマーカジェネレータは、自動的に制御またはユーザ操作され得るＣＣＵ　１０１によって制御され得る。特定の実施形態において、バイオマーカジェネレータ１０２は、データをトレーニングデータセットとテストデータセットとにランダムに分割する度に、複数の候補バイオマーカを生成するように動作し得る。そのような複数の候補バイオマーカを生成するために、バイオマーカジェネレータ１０２の動作は、複数回、反復され得る。ＣＣＵ　１０１は、所望の数の候補バイオマーカを含む１つ以上のシステム反復パラメータを受信し得、それらは、次に、バイオマーカジェネレータ１０２の動作が反復され得る回数を決定するように使用され得る。ＣＣＵ　１０１はまた、バイオマーカ中の構成要素の数（例えば、バイオマーカ遺伝子シグネチャ中の遺伝子の数）を表し得る所望のバイオマーカサイズを含む他のシステムパラメータを受信し得る。バイオマーカサイズ情報は、トレーニングデータから候補バイオマーカを生成するために分類器性能監視エンジン１１６によって使用され得る。バイオマーカジェネレータ１０２の動作、特定すると分類エンジン１１４の動作は、図２～４への参照によってさらに詳細に説明される。

【００２０】

バイオマーカジェネレータ１０２は、１つ以上の候補バイオマーカおよび候補エラー率を生成し、それらは、ロバストなバイオマーカを生成するためにバイオマーカコンソリデータ１０４によって使用される。バイオマーカコンソリデータ１０４は、複数の候補バイオマーカを受信して複数の候補バイオマーカにわたって最も頻繁に発生する遺伝子を有する新しいバイオマーカシグネチャを生成するバイオマーカコンセンサスエンジン１２８を含む。バイオマーカコンソリデータ１０４は、複数の候補バイオマーカにわたって全体的なエラー率を決定するためのエラー計算エンジン１３０を含む。バイオマーカジェネレータ１０２と同様に、バイオマーカコンソリデータ１０４もまた、自動的に制御またはユーザ操作され得るＣＣＵ　１０１によって制御され得る。ＣＣＵ　１０１は、最小バイオマーカサイズについての好適な閾値を受信および／または決定し得、バイオマーカジェネレータ１０２およびバイオマーカコンソリデータ１０４の両方を動作させる反復の数を決定

するように、この情報を使用し得る。１つの実施形態において、各反復中に、ＣＣＵ　１０１は、バイオマーカサイズを１つ減少させ、閾値が達せられるまでバイオマーカジェネレータ１０２およびバイオマーカコンソリデータ１０４の両方を反復する。そのような実施形態において、バイオマーカコンセンサスエンジン１２８は、各反復について、新しいバイオマーカシグネチャおよび新しい全体的なエラー率を出力する。したがって、バイオマーカコンセンサスエンジン１２８は、閾値から最大バイオマーカサイズまで様々である異なるサイズを各々が有する一組の新しいバイオマーカシグネチャ（複数）を出力する。バイオマーカコンソリデータ１０４はさらに、これらの新しいバイオマーカシグネチャの各々の性能尺度またはエラー率を検討して出力のために最適なバイオマーカを選択するバイオマーカ選択エンジン１２６を含む。バイオマーカコンソリデータ１０４およびそれぞれのエンジンの動作は、図２～４への参照によってさらに詳細に説明される。

【００２１】

図３は、データセットを分類するための例示的なプロセスの流れ図である。ステップ３０２で、分類エンジン１１４は、トレーニングデータおよびテストデータを受信する。下記で説明されるように、分類エンジン１１４は、１つ以上の分類器を開発するためにトレーニングデータを使用し、次いで、１つ以上の分類器をテストデータに適用する。図３で図示されるように、トレーニングデータは、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４と、トレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６とを含む。トレーニングデータセットＴ０．ｔｒａｉｎ　３０４中の各要素は、データサンプル（例えば、特定の患者からの発現データのベクトル）を表し、トレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６中の既知のクラス識別子に対応する。例えば、３クラスシナリオにおいて、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４中の第１の要素は、特定の疾患を有する患者についての遺伝子発現データを表し得、トレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６中の第１の要素「疾患陽性」に対応し得、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４中の第２の要素は、特定の疾患に耐性または免疫がある患者についての遺伝子発現データを表し得、トレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６中の第２の要素「疾患免疫性」に対応し得、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４中の第３の要素は、特定の疾患がない患者についての遺伝子発現データを表し得、トレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６中の第３の要素「疾患なし」に対応し得る。ステップ３０２で受信されるテストデータは、テストデータセットＴ０．ｔｅｓｔ　３０８を含み、そのテストデータセットＴ０．ｔｅｓｔは、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４中のデータサンプルと同一の基礎的な種類のデータを表すが、例えば、異なる患者または異なる実験から採取されたサンプルを表し得る。任意選択で、分類エンジン１１４はまた、分類器がテストデータセットＴ０．ｔｅｓｔ　３０８に適用される場合に分類エンジン１１４によって生成される分類器の性能を評価するために使用され得る、テストデータセット中のデータサンプルについての既知のクラス識別子を含むテストクラスセットｃｌ．ｔｅｓｔ　３１０を受信する。いくつかの実装において、テストデータセットＴ０．ｔｅｓｔ　３０８中のデータサンプルについてのいかなる既知のクラスも利用可能ではなく、したがって、テストクラスセットｃｌ．ｔｅｓｔ　３１０は、分類エンジン１１４に提供されない。

【００２２】

概して、ステップ３０２で受信されるデータは、サンプル中の複数の異なる遺伝子の発現値等の、分類が引き出され得る任意の実験データまたは別様に得られたデータ、および／または、任意の生物学的に意味のある被分析物のレベル等の種々の表現型の特性を表し得る。特定の実施形態において、データセットは、疾患状態についてのおよび対照状態についての発現レベルデータを含み得る。本明細書で使用される場合、「遺伝子発現レベル」という用語は、遺伝子によってコード化される分子（例えば、ＲＮＡまたはポリペプチド）の量を指し得る。ｍＲＮＡ分子の発現レベルは、ｍＲＮＡの量（ｍＲＮＡをコード化する遺伝子の転写活性によって決定される）、および、ｍＲＮＡの安定性（ｍＲＮＡの半減期によって決定される）を含み得る。遺伝子発現レベルはまた、遺伝子によってコード

化される所与のアミノ酸配列に対応するポリペプチドの量を含み得る。したがって、遺伝子の発現レベルは、遺伝子から転写されるｍＲＮＡの量、遺伝子によってコード化されるポリペプチドの量、または、それら両方に対応することができる。遺伝子の発現レベルはさらに、遺伝子産物の異なる形態の発現レベルによってカテゴライズされ得る。例えば、遺伝子によってコード化されるＲＮＡ分子は、差次的に発現させられたスプライスバリアント（ｄｉｆｆｅｒｅｎｔｉａｌｌｙ　ｅｘｐｒｅｓｓｅｄ　ｓｐｌｉｃｅ　ｖａｒｉａｎｔ）、異なる開始または終結部位を有する転写産物、および／または、他の特異的に処理された形態を含み得る。遺伝子によってコード化されるポリペプチドは、ポリペプチドの開裂および／または修飾形態を含み得る。ポリペプチドは、リン酸化、脂質化、プレニル化、硫酸化、水酸化、アセチル化、リボシル化、ファルネシル化、炭水化物の追加、および、同等物によって修飾されることができる。さらに、所与の種類の修飾を有するポリペプチドの複数の形態が、存在し得る。例えば、ポリペプチドは、複数の部位においてリン酸化され、異なるレベルの特異的にリン酸化されたタンパク質を発現し得る。

【００２３】

特定の実施形態において、細胞または組織における遺伝子発現レベルは、遺伝子発現プロファイルによって表され得る。遺伝子発現プロファイルは、細胞または組織等の検体における遺伝子の発現レベルの特徴的な表現を指し得る。個体からの検体における遺伝子発現プロファイルの決定は、個体の遺伝子発現状態を表す。遺伝子発現プロファイルは、メッセンジャーＲＮＡまたはポリペプチドの発現、あるいは、細胞中または組織中の１つ以上の遺伝子によってコード化されるそれらの形態を反映する。発現プロファイルは、概して、異なる細胞または組織の間で異なる発現パターンを示す生体分子（核酸、タンパク質、炭水化物）のプロファイルを指し得る。遺伝子発現プロファイルを表すデータサンプルは、発現レベルのベクトルとして記憶され得、ベクトルにおける各入力は、特定の生体分子または他の生物学的実体に対応する。

【００２４】

特定の実施形態において、データセットは、サンプル中の複数の異なる遺伝子の遺伝子発現値を表す要素を含み得る。他の実施形態において、データセットは、質量分析によって検出されるピークを表す要素を含み得る。概して、各データセットは、複数の生物学的状態クラスのうちの１つに各々が対応するデータサンプル（複数）を含み得る。例えば、生物学的状態クラスは、サンプルのソース（すなわち、サンプルが取得される患者）における疾患の有無、病期、疾患のリスク、疾患の再発の可能性、１つ以上の遺伝子座における共有遺伝子型（例えば、共通ＨＬＡハプロタイプ、遺伝子における突然変異、メチル化等の遺伝子の修飾等）、作用物質（例えば、毒性物質または潜在的に毒性の物質、環境汚染物質、候補薬剤等）または条件（温度、ｐＨ等）への曝露、人口学的特性（年齢、性別、体重、家族歴、既往歴等）、作用物質への耐性、作用物質への感受性（例えば、薬剤への反応性）、および、同等物を含むことができるが、それらに限定されない。

【００２５】

データセットは、最終的な分類器選択における収集バイアスを低減するように、互いから独立し得る。例えば、それらは、複数のソースから収集されることができ、異なる除外または包含の基準を使用して異なる時間に異なる場所から収集され得、すなわち、データセットは、生物学的状態クラスを定義する特性外の特性を考慮する場合に、比較的ヘテロジニアスであり得る。ヘテロジェナイティ（ｈｅｔｅｒｏｇｅｎｅｉｔｙ）に寄与する要因は、性別、年齢、民族性による生物学的変動、摂食、運動、睡眠の挙動による個体的変動、および、血液処理のための臨床プロトコルによるサンプル取り扱い変動を含むが、それらに限定されない。しかしながら、生物学的状態クラスは、１つ以上の共通特性を備え得る（例えば、サンプルソースは、疾患および同一の性別、または、１つ以上の他の共通の人口学的特性を有する個体を表し得る）。特定の実施形態において、複数のソースからのデータセットは、異なる時間および／または異なる条件下における患者の同一の集団からのサンプルの収集によって生成される。

【００２６】

特定の実施形態において、複数のデータセットは、複数の異なる臨床試験場から取得され、各データセットは、各個別試験場で取得される複数の患者サンプルを備える。サンプル種類は、血液、血清、血漿、乳頭吸引物、尿、涙、唾液、髄液、リンパ液、細胞および／または組織溶解物、レーザ顕微解剖組織または細胞サンプル、（例えば、パラフィンブロック中の、または、凍結された）埋め込み細胞または組織、（例えば、剖検からの）新鮮なまたは保存用のサンプルを含むが、それらに限定されない。サンプルは、例えば、インビトロで細胞または組織培養から得ることができる。代替として、サンプルは、生体から、または、単細胞生物等の生物の集団から得ることができる。１つの例において、特定の癌についてのバイオマーカを識別する場合、２つのテスト場で独立したグループによって選択される対照から、血液サンプルが収集され、それによって、独立した独立したデータセットが開発されるサンプルを提供し得る。

【００２７】

いくつかの実装において、トレーニングセットおよびテストセットは、バルクデータを受信してそのバルクデータをトレーニングデータセットとテストデータセットとに分割するデータ前処理エンジン１１０（図１）によって生成される。特定の実施形態において、データ前処理エンジン１１０は、データをこれら２つのグループにランダムに分割する。データをランダムに分割することが、クラスを予測してロバストな遺伝子シグネチャを生成するために望ましくあり得る。他の実施形態において、データ前処理エンジン１１０は、データの種類または標識に基づいて、データを２つ以上のグループに分割する。概して、データは、本開示の範囲から逸脱することなく、所望に応じた任意の好適な方法で、トレーニングデータセットおよびテストデータセットに分割されることができる。トレーニングデータセットおよびテストデータセットは、任意の好適なサイズを有し得、同一のまたは異なるサイズであり得る。特定の実施形態において、データ前処理エンジン１１０は、データをトレーニングデータセットとテストデータセットとに分割することの前に、１つ以上のデータを破棄し得る。特定の実施形態において、データ前処理エンジン１１０は、任意のさらなる処理の前に、トレーニングデータセットおよび／またはテストデータセットから１つ以上のデータを破棄し得る。

【００２８】

ステップ３１１において、分類エンジン１１４は、カウンタ変数ｉを１に等しく設定する。ステップ３１２において、分類エンジン１１４は、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４およびトレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６に基づいて、第１の分類器ｒｆ　３１４を生成する。図２は、データセット中の要素の分類を図示する。分類エンジン１１４は、サポートベクトルマシン技法、線形判別分析技法、ランダムフォレスト技法、ｋ最近傍技法、部分最小二乗技法（部分最小二乗および線形判別分析特徴を組み合わせる技法を含む）、ロジスティック回帰技法、ニューラルネットワークベースの技法、決定木ベースの技法、および、（例えば、「Ｄｉａｇｎｏｓｉｓ　ｏｆ　ｍｕｌｔｉｐｌｅ　ｃａｎｃｅｒ　ｔｙｐｅｓ　ｂｙ　ｓｈｒｕｎｋｅｎ　ｃｅｎｔｒｏｉｄｓ　ｏｆ　ｇｅｎｅ　ｅｘｐｒｅｓｓｉｏｎ」ＰＮＡＳ，ｖ．９９，ｎ．１０，２００２で、Ｔｉｂｓｈｉｒａｎｉ、Ｈａｓｔｌｅ、Ｎａｒａｓｉｍｈａｎ、および、Ｃｈｕによって説明されるような）収縮重心技法（ｓｈｒｕｎｋｅｎ　ｃｅｎｔｒｏｉｄ　ｔｅｃｈｎｉｑｕｅ）を含むが、それらに限定されないいずれか１つ以上の既知の機械学習アルゴリズムをステップ３１２で使用し得る。いくつかのそのような技法は、線形判別分析、サポートベクトルマシン、ランダムフォレスト（Ｂｒｅｉｍａｎ，Ｍａｃｈｉｎｅ　Ｌｅａｒｎｉｎｇ，４５（１）：５－３２　（２００１））、ｋ最近傍（Ｂｉｓｈｏｐ，Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋｓ　ｆｏｒ　Ｐａｔｔｅｒｎ　Ｒｅｃｏｇｎｉｔｉｏｎ，ｅｄ．Ｏ．Ｕ．Ｐｒｅｓｓ，１９９５）、部分最小二乗判別分析、および、ＰＡＭＲ（Ｔｉｂｓｈｉｒａｎｉ　ｅｔ　ａｌ．，Ｐｒｏｃ　Ｎａｔｌ　Ａｃａｄ　Ｓｃｉ　ＵＳＡ，９９（１０）：６５６７－６５７２　（２００２））に対応する、ｌｄａ、ｓｖｍ、ｒａｎｄｏｍＦｏｒｅｓｔ、ｋｎｎ、ｐｌｓ．ｌｄａ、および、ｐａｍｒを含むＲプログラミング言語用パッケージとして利用可能である。分類エンジ

ン１１４は、ステップ３１２で、第１の分類器ｒｆ　３１４をメモリに記憶し得る。
【００２９】
ステップ３１６において、分類エンジン１１４は、第１の分類器ｒｆ　３１４（ステップ３１２で生成される）をテストデータセットＴ０．ｔｅｓｔ　３０８に適用することによって、一組の予測されたテスト分類ｐｒｅｄｃｌ．ｔｅｓｔ　３１８を生成する。分類エンジン１１４は、ステップ３１６で、予測された分類ｐｒｅｄｃｌ．ｔｅｓｔ　３１８をメモリに記憶し得る。
【００３０】
ステップ３２０において、分類エンジン１１４は、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４を変換する。この変換は、トレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６に基づいてトレーニングデータセットＴ０．ｔｒａｉｎ　３０４を変換する変換関数ｃｏｒｒｅｃｔｅｄＤａｔａに従って進む。ステップ３１０の変換の結果は、分類エンジン１１４がメモリに記憶し得る変換されたトレーニングデータセットＴ０．ｔｒａｉｎ．２　３２２である。いくつかの実装において、ステップ３２０で分類エンジン１１４によって行われる変換は、バイアス補正技法を含む。例えば、変換は、全体として採取されるデータセットの重心、または、データセットにおいて表される各クラスの重心に関して、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４の要素を調整することによって、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４を「再び中心に置いて」もよい。
【００３１】
１つの特定の再中心化技法は、異なるグループの重心の中心に基づいて、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４の要素を中心に置くことを伴う。トレーニングデータセットＴ０．ｔｒａｉｎ　３０４中にｎ個のデータサンプルがあり、かつ、各データサンプルがｐ個の入力（例えば、ｐ個の異なる遺伝子ついての発現レベルを表す）を有するベクトルである場合、ｘｉｊにデータサンプルｊのｉ番目の入力を表させる。トレーニングクラスセットｃｌ．ｔｒａｉｎ　３０８がＫ個の異なるクラスを表す場合、クラスｋにおけるｎｋ個のサンプルの指数をＣｋに表させる。分類エンジン１１４は、クラスｋの重心のｉ番目の成分を下記のように計算し得、
【数１】

$$\overline{x}_{ik} = \sum_{j \in C_k} \frac{x_{ij}}{n_k} \quad . \qquad (1)$$

かつ、クラス重心の中心のｉ番目の成分を下記のように計算し得る。
【数２】

$$\overline{x}_i^c = \sum_{k=1}^{K} \frac{\overline{x}_{ik}}{K} \quad . \qquad (2)$$

【００３２】
分類エンジン１１４はまた、全体的な重心のｉ番目の成分を下記のように計算し得る。
【数３】

$$\overline{x}_i = \sum_{j=1}^{n} \frac{x_{ij}}{n} \quad . \qquad (3)$$

【００３３】
次いで、分類エンジン１１４は、下記によって求められる差を加えることによって、ト

レーニングデータセットＴ０．ｔｒａｉｎ　３０４の各要素の中のｉ番目の入力を調整することを含む変換を行ってもよい。

【数４】

$$\Delta = -\overline{x}_i^c \quad . \qquad (4)$$

【００３４】

いくつかの実装において、ステップ３２０で行われる変換は、方程式１～４への参照によって上記で説明されるもの以外の偏移（ｓｈｉｆｔ）、回転、シアー、これらの変換の組み合わせ、または、任意の他の線形あるいは非線形の変換を含む。

【００３５】

ステップ３２４において、分類エンジン１１４は、テストデータセットＴ０．ｔｅｓｔ　３０８を変換する。テストデータセットＴ０．ｔｅｓｔ　３０８に適用される変換、ｃｏｒｒｅｃｔｅｄＤａｔａは、ステップ３２０でトレーニングデータセットＴ０．ｔｒａｉｎ　３０４に適用される同一の種類の変換であるが、Ｔ０．ｔｒａｉｎ　３０４およびｐｒｅｄｃｌ．ｔｒａｉｎ　３１４の代わりに、引数Ｔ０．ｔｅｓｔ　３０８およびｐｒｅｄｃｌ．ｔｅｓｔ　３１８に関して適用される。例えば、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４の要素が、トレーニングデータセットＴ０．ｔｒａｉｎ　３０４のクラスの重心に関して計算されるような方程式４によって求められるΔの値によって、ステップ３２０で調整される場合には、テストデータセットＴ０．ｔｅｓｔ　３０８の要素は、テストデータセットＴ０．ｔｅｓｔ　３０８のクラスの重心に関して計算されるような方程式４によって求められるΔの値によって、ステップ３２４で調整される。ステップ３２４の変換の結果は、分類エンジン１１４がメモリに記憶し得る変換されたテストデータセットＴ０．ｔｅｓｔ．２　３２６である。

【００３６】

ステップ３２７において、分類エンジン１１４は、反復カウンタｉの値が１に等しいかどうかを決定する。そうである場合、分類エンジン１１４は、分類エンジン１１４が、第２の分類器ｒｆ２　３２９を生成するために、変換されたトレーニングデータセットＴ０．ｔｒａｉｎ．２　３２２およびトレーニングクラスセットｃｌ．ｔｒａｉｎ　３０６を使用するステップ３２８を続けて実行する。ステップ３３２およびステップ３３６への参照によって上記で説明されるように、任意の機械学習技法が、ステップ３２８で分類器を生成するために適用され得る。第２の分類器ｒｆ２　３２９は、第１の分類器ｒｆ　３１４（例えば、両方のＳＶＭ分類器）と同一の種類であり得るか、または、異なる種類であり得る。

【００３７】

ステップ３３１において、分類エンジン１１４は、反復カウンタｉをインクリメントし、次いで、分類エンジン１１４が第２の分類器ｒｆ２　３２９を（ステップ３２４で分類エンジン１１４によって生成されるような）変換されたテストデータセットＴ０．ｔｅｓｔ．２　３２６に適用するステップ３３３を続けて実行する。ステップ３３３の出力は、変換されたデータセットＴ０．ｔｅｓｔ．２　３２６のための一組の予測された分類ｐｒｅｄｃｌ．ｔｅｓｔ．２　３３０である。分類エンジン１１４は、表示デバイス、印刷デバイス、記憶デバイス、ネットワークにわたって分類エンジン１１４と通信している別のデバイス、または、システム１００の内部あるいは外部の任意の他のデバイスに、予測された分類を出力し得る。

【００３８】

ステップ３３２において、分類エンジン１１４は、（ステップ３１６で生成されるような）予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ　３１８の分類と（ステップ３２８で生成されるような）予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ．２　３３０の分類との間に何らかの差異があるかどうかを決定する。予測された分類のセットが一致する（すなわち、テストデータセットＴ０．ｔｅｓｔ　３０８中の各データサンプルについて、そ

のデータサンプルについての予測されたクラスが、２つの予測された分類セットの間で同一である）場合には、分類エンジン１１４は、ステップ３３８へ進み、予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ．２　３３０（同等に、予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ　３１８）をテストデータセットＴ０．ｔｅｓｔ　３０８の最終的な分類として出力する。

【００３９】

分類エンジン１１４が分類データセットｐｒｅｄｃｌ．ｔｅｓｔ　３１８と分類データセットｐｒｅｄｃｌ．ｔｅｓｔ．２　３３０との間の差異を識別する場合、分類エンジン１１４は、ステップ３３４へ進み、テストデータセットＴ０．ｔｅｓｔ　３０８の以前に記憶された値を、（ステップ３２４の変換によって生成されるような）変換されたテストデータセットＴ０．ｔｅｓｔ．２　３２６の値と置換する。結果として、テストデータセットＴ０．ｔｅｓｔ　３０８は、変換されたテストデータセットＴ０．ｔｅｓｔ．２　３２６の値を有する。分類エンジン１１４は、ステップ３３６へ進み、（ステップ３１６で生成されるような）予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ　３１８の以前に記憶された値を、（ステップ３２８で生成されるような）予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ．２　３３０の値と置換する。結果として、予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ　３１８は、予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ．２　３３０の値を有する。

【００４０】

テストデータセットＴ０．ｔｅｓｔ　３０８の値が変換されたテストデータセットＴ０．ｔｅｓｔ．２　３２６の値で更新され、かつ、予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ　３１８が予測された分類セットｐｒｅｄｃｌ．ｔｅｓｔ．２　３３０の値で更新されると、分類エンジン１１４は、ステップ３２４に戻って新しい変換を行い、分類エンジン１１４が（ステップ３３２で）予測された分類の間に差異がないことを決定するまで、このプロセスを反復する。

【００４１】

分類器性能監視エンジン１１６は、好適な性能測定基準を使用して、図３のプロセスの終わりに、分類エンジン１１４によって生成される最終的な分類の性能を分析し得る。特定の実施形態において、性能測定基準は、エラー率を含み得る。性能測定基準はまた、試行された予測の総数によって除算された正しい予測の数を含み得る。性能測定基準は、本開示の範囲から逸脱することなく、任意の好適な尺度であり得る。

【００４２】

本主題の実装は、本明細書で説明されるような１つ以上の特徴と、１つ以上の機械（例えば、コンピュータ、ロボット）に本明細書で説明される動作を実現させるように動作可能な機械可読媒体を備える物品とを備えるシステム、方法、および、コンピュータプログラム製品を含むことができるが、それらに限定されない。本明細書で説明される方法は、単一のコンピューティングシステムまたは複数のコンピューティングシステムに存在する１つ以上のプロセッサまたはエンジンによって実装されることができる。そのような複数のコンピューティングシステムは、接続されることができ、複数のコンピューティングシステムのうちの１つ以上の間の直接接続を介したネットワーク（例えば、インターネット、無線広域ネットワーク、ローカルエリアネットワーク、広域ネットワーク、有線ネットワーク、または、同等物）を経由した接続を含むが、それに限定されない１つ以上の接続を介して、データおよび／またはコマンド、あるいは、他の命令または同等物を交換することができる。

【００４３】

図４は、図１～３への参照によって説明されるプロセスを行うための回路を含む図１のシステム１００の構成要素のうちのいずれか等の、コンピューティングデバイスのブロック図である。システム１００の構成要素の各々は、１つ以上のコンピューティングデバイス４００上に実装され得る。特定の局面において、複数の上記の構成要素およびデータベースは、１つのコンピューティングデバイス４００内に含まれ得る。特定の実装において

、構成要素およびデータベースは、いくつかのコンピューティングデバイス４００にわたって実装され得る。

【００４４】

コンピューティングデバイス４００は、少なくとも１つの通信インターフェースユニットと、入力／出力コントローラ４１０と、システムメモリと、１つ以上のデータ記憶デバイスとを備える。システムメモリは、少なくとも１つのランダムアクセスメモリ（ＲＡＭ　４０２）と、少なくとも１つの読み取り専用メモリ（ＲＯＭ　４０４）とを含む。これらの要素は全て、中央処理ユニット（ＣＰＵ　４０６）と通信し、コンピューティングデバイス４００の動作を促進する。コンピューティングデバイス４００は、多くの異なる方法で構成され得る。例えば、コンピューティングデバイス４００は、従来のスタンドアロンコンピュータであり得るか、または、代替として、コンピューティングデバイス４００の機能は、複数のコンピュータシステムおよびアーキテクチャにわたって分散され得る。コンピューティングデバイス４００は、データ分割、区別、分類、スコア化、ランク付け、および、記憶の動作のうちのいくつかまたは全てを行うように構成され得る。図４において、コンピューティングデバイス４００は、ネットワークまたはローカルネットワークを介して、他のサーバまたはシステムにリンクされる。

【００４５】

コンピューティングデバイス４００は、分散されたアーキテクチャにおいて構成され得、データベースおよびプロセッサは、別個のユニットまたは場所において格納される。いくつかのそのようなユニットは、一次処理機能を行い、最低限でも、一般コントローラまたはプロセッサおよびシステムメモリを含む。そのような局面において、これらのユニットの各々は、通信インターフェースユニット４０８を介して、他のサーバ、クライアント、または、ユーザコンピュータ、および、他の関連デバイスとの一次通信リンクとしての役割を果たす通信ハブまたはポート（図示せず）に取り付けられる。通信ハブまたはポートは、それ自体が最小処理能力を有し、主に、通信ルータとしての役割を果たし得る。種々の通信プロトコルは、限定されないが、Ｅｔｈｅｒｎｅｔ（登録商標）、ＳＡＰ、ＳＡＳ（登録商標）、ＡＴＰ、Ｂｌｕｅｔｏｏｔｈ（登録商標）、ＧＳＭ（登録商標）、および、ＴＣＰ／ＩＰを含むシステムの一部であり得る。

【００４６】

ＣＰＵ　４０６は、１つ以上の従来のマイクロプロセッサ等のプロセッサ、および、ＣＰＵ　４０６から作業負荷をオフロードするための数値演算コプロセッサ等の１つ以上の補助コプロセッサを備える。ＣＰＵ　４０６は、通信インターフェースユニット４０８および入力／出力コントローラ４１０と通信し、それらを通して、ＣＰＵ　４０６は、他のサーバ、ユーザ端末、または、デバイス等の他のデバイスと通信する。通信インターフェースユニット４０８および入力／出力コントローラ４１０は、例えば、他のプロセッサ、サーバ、または、クライアント端末と同時に通信するための複数の通信チャネルを含み得る。相互に通信しているデバイスは、継続的に相互に伝送している必要はない。反対に、そのようなデバイスは、必要に応じて相互に伝送する必要しかなく、実際には、ほとんどの時間、データを交換することを控え得、いくつかのステップが行われることを要求することにより、デバイス間の通信リンクを確立し得る。

【００４７】

ＣＰＵ　４０６はまた、データ記憶デバイスと通信する。データ記憶デバイスは、磁気、光学、または、半導体のメモリの適切な組み合わせを備え得、例えば、ＲＡＭ　４０２、ＲＯＭ　４０４、フラッシュドライブ、コンパクトディスクまたはハードディスクあるいはドライブ等の光学ディスクを含み得る。ＣＰＵ　４０６およびデータ記憶デバイスは、各々、例えば、単一のコンピュータまたは他のコンピューティングデバイス内に全体的に位置し得るか、または、ＵＳＢポート、シリアルポートケーブル、同軸ケーブル、Ｅｔｈｅｒｎｅｔ（登録商標）型ケーブル、電話回線、無線周波数送受信機、または、他の類似の無線もしくは有線の媒体、あるいは、前述のものの組み合わせ等の通信媒体によって、相互に接続され得る。例えば、ＣＰＵ　４０６は、通信インターフェースユニット４０

８を介して、データ記憶デバイスに接続され得る。ＣＰＵ　４０６は、１つ以上の特定の処理機能を行なうように構成され得る。

【００４８】

データ記憶デバイスは、例えば、（ｉ）コンピューティングデバイス４００のためのオペレーティングシステム４１２、（ｉｉ）本明細書で説明されるシステムおよび方法に従って、特に、ＣＰＵ　４０６に関して詳細に説明されるプロセスに従って、ＣＰＵ　４０６に命令するように適合させられた１つ以上のアプリケーション４１４（例えば、コンピュータプログラムコードまたはコンピュータプログラム製品）、または、（ｉｉｉ）プログラムによって要求される情報を記憶するために利用され得る情報を記憶するように適合させられたデータベース（単数または複数）４１６を記憶し得る。いくつかの局面において、データベースは、実験データ、および、既刊文献モデルを記憶するデータベース（単数または複数）を含む。

【００４９】

オペレーティングシステム４１２およびアプリケーション４１４は、例えば、圧縮、アンコンパイル、および、暗号化されたフォーマットにおいて記憶され得、コンピュータプログラムコードを含み得る。プログラムの命令は、ＲＯＭ　４０４またはＲＡＭ　４０２から等、データ記憶デバイス以外のコンピュータ可読媒体から、プロセッサのメインメモリに読み込まれ得る。プログラムにおける命令のシーケンスの実行は、ＣＰＵ　４０６に、本明細書で説明されるプロセスステップを行なわせるが、有線回路が、本発明のプロセスの実装のためのソフトウェア命令の代わりに、または、それと組み合わせて使用され得る。したがって、説明されるシステムおよび方法は、ハードウェアおよびソフトウェアの任意の特定の組み合わせに限定されない。

【００５０】

好適なコンピュータプログラムコードは、本明細書で説明されるようなモデル化、スコア化、および、集約に関連する１つ以上の機能を果たすために提供され得る。プログラムはまた、オペレーティングシステム４１２、データベース管理システム、および、プロセッサが入力／出力コントローラ４１０を介してコンピュータ周辺デバイス（例えば、ビデオディスプレイ、キーボード、コンピュータマウス等）と連動することを可能にする「デバイスドライバ」等のプログラム要素を含み得る。

【００５１】

コンピュータ可読命令を備えるコンピュータプログラム製品も、提供される。コンピュータ可読命令は、コンピュータシステム上にロードされて実行される場合、本方法または上記で説明される方法の１つ以上のステップに従って、コンピュータシステムを動作させる。本明細書で使用される場合、「コンピュータ可読媒体」という用語は、実行のために、コンピューティングデバイス４００のプロセッサ（または、本明細書で説明されるデバイスの任意の他のプロセッサ）に命令を提供するかまたは提供に関与する任意の非一時的媒体を指す。そのような媒体は、不揮発性媒体および揮発性媒体を含むが、それらに限定されない多くの形態をとり得る。不揮発性媒体は、例えば、光学、磁気、または、光磁気のディスク、あるいは、フラッシュメモリ等の集積回路メモリを含む。揮発性媒体は、典型的にメインメモリを構成するダイナミックランダムアクセスメモリ（ＤＲＡＭ）を含む。コンピュータ可読媒体の共通の形態は、例えば、フロッピー（登録商標）ディスク、フレキシブルディスク、ハードディスク、磁気テープ、任意の他の磁気媒体、ＣＤ－ＲＯＭ、ＤＶＤ、任意の他の光学媒体、パンチカード、ペーパーテープ、孔のパターンを有する任意の他の物理的媒体、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、または、ＥＥＰＲＯＭ（電気的に消去可能なプログラマブル読み取り専用メモリ）、ＦＬＡＳＨ－ＥＥＰＲＯＭ、任意の他のメモリチップまたはカートリッジ、あるいは、コンピュータが読み取ることができる任意の他の非一時的媒体を含む。

【００５２】

コンピュータ可読媒体の種々の形態は、実行のために、１つ以上の命令の１つ以上のシーケンスをＣＰＵ　４０６（または本明細書で説明されるデバイスの任意の他のプロセッ

サ）に搬送することに関与し得る。例えば、命令は、最初に、遠隔コンピュータ（図示せず）の磁気ディスク上にあり得る。遠隔コンピュータは、命令をその動的メモリ内にロードし、Ｅｔｈｅｒｎｅｔ（登録商標）接続、ケーブルライン、または、モデムを使用する電話回線をも経由して、命令を送信することができる。コンピューティングデバイス４００（例えば、サーバ）にローカルの通信デバイスは、それぞれの通信ライン上でデータを受信し、プロセッサのためのシステムバス上にデータを置くことができる。システムバスは、データをメインメモリに搬送し、そこから、プロセッサは、命令を読み出して実行する。メインメモリによって受信される命令は、任意選択で、プロセッサによる実行の前または後のいずれかにおいて、メモリに記憶され得る。加えて、命令は、通信ポートを介して、種々のタイプの情報を搬送する無線通信またはデータストリームの例示的形態である電気的、電磁的、または、光学的な信号として受信され得る。

【実施例】

【００５３】

下記の公開データセットを、Ｇｅｎｅ　Ｅｘｐｒｅｓｓｉｏｎ　Ｏｍｎｉｂｕｓ（ＧＥＯ）（ｈｔｔｐ：／／ｗｗｗ．ｎｃｂｉ．ｎｌｍ．ｎｉｈ．ｇｏｖ／ｇｅｏ／）リポジトリからダウンロードする。

【表１】

a. GSE10106 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10106)
b. GSE10135 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10135)
c. GSE11906 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11906)
d. GSE11952 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11952)
e. GSE13933 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13933)
f. GSE19407 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19407)
g. GSE19667 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19667)
h. GSE20257 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20257)
i. GSE5058 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5058)
j. GSE7832 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7832)
k. GSE8545 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8545).

【００５４】

トレーニングデータセットは、Ａｆｆｙｍｅｔｒｉｘプラットフォーム（ＨＧＵ－１３３＋２）上にある。未加工データファイルを、Ｒ（Ｒ　Ｄｅｖｅｌｏｐｍｅｎｔ　Ｃｏｒｅ　Ｔｅａｍ，　２００７）の中のＢｉｏｃｏｎｄｕｃｔｏｒ（Ｇｅｎｔｌｅｍａｎ，２００４）に属するａｆｆｙパッケージ（Ｇａｕｔｉｅｒ，　２００４）のＲｅａｄＡｆｆｙ機能によって読み取り、品質を、ＲＮＡ分解プロット（ａｆｆｙパッケージのＡｆｆｙＲＮＡｄｅｇ機能を伴う）、ＮＵＳＥ、および、ＲＬＥプロット（機能ａｆｆｙＰＬＭ（Ｂｒｅｔｔｓｃｈｎｅｉｄｅｒ，　２００８）を伴う）を生成し、ＭＡ（ＲＬＥ）値を計算し、品質管理チェック上の一組の閾値を下回るか、または、上記のデータセットの中で複製されるトレーニングデータセットからアレイを除外し、ｇｃｒｍａアルゴリズム（Ｗｕ，　２００４）を使用して品質管理チェックに合格するアレイを正規化することによって、管理する。トレーニングセットサンプル分類を、各データセットについてのＧＥＯデータベースのシリーズマトリクスファイルから取得する。出力は、２３３個のサンプル（２８個のＣＯＰＤサンプルおよび２０５個の対照サンプル）についての５４６７５個のプローブセットを伴う遺伝子発現マトリクスから成る。均衡の取れたデータセットを作製するために、ＣＯＰＤサンプルは、同時係属中の米国仮特許出願第６１／６６２８１２号で説明されるようなＤｕａｌ　Ｅｎｓｅｍｂｌｅ方法を適用する前に、２２４個のＣＯＰＤサンプルを取得するための多重時間（ｍｕｌｔｉｐｌｅ　ｔｉｍｅ）であった。２０５

人の対照および２２４人のＣＯＰＤ患者を含む複合データセットを用いて、４０９個の遺伝子を有する遺伝子シグネチャを構築した。８５０個の二進値を、ランダムベクトルにおいて使用した。本方法で使用される分類方法は、下記のＲパッケージ、すなわち、ｌｄａ、ｓｖｍ、ｒａｎｄｏｍＦｏｒｅｓｔ、ｋｎｎ、ｐｌｓ．ｌｄａ、および、ｐａｍｒを含んでいた。最大反復を、５０００であるように設定した。マシューズ相関係数（ＭＣＣ）、トレーニングデータセットにおける相互検証プロセスの精度は、それぞれ、０．７４３、０．８７である。トレーニングデータセット中の遺伝子シグネチャのヒートマップを、図５に示す。図５のヒートマップにおいて、遺伝子発現値を、行ごとに中心に置いた。ヒートマップの色は、グレースケールでは明確に示されない場合もあるが、図５のデータは、対照データが左に示され、ＣＯＰＤデータが右側に示されていることを示す。テストデータセットは、１６個の対照サンプルおよび２４個のＣＯＰＤサンプルを含む民間供給業者（Ｇｅｎｅｌｏｇｉｃ）から入手した未公開データセットである。本発明の変換不変方法を適用することなく、Ｄｕａｌ　Ｅｎｓｅｍｂｌｅによって生成される遺伝子シグネチャは、合計４０個のサンプルうちの２９個のサンプルを正しく予測した。精度は０．７２５であり、ＭＣＣは０．５２７である。１６個の対照サンプルにおいて、遺伝子シグネチャは、１５個を対照として正しく予測したが、１個をＣＯＰＤとして誤って予測した。２４個のＣＯＰＤサンプルの間で、遺伝子シグネチャは、１４個をＣＯＰＤサンプルとして正しく予測したが、１０個を対照として誤って予測した。

【００５５】

しかしながら、変換不変方法が適用された場合、２つまたは複数のクラスの中心、および、１００に設定された最大反復に従って偏移（ｓｈｉｆｔ）を伴った。同一の遺伝子シグネチャは、合計４０個のサンプルのうちの３０個のサンプルを正しく予測した。精度は０．７５であり、ＭＣＣは０．５３３である。１６個の対照サンプルにおいて、遺伝子シグネチャは、１４個を対照として正しく予測したが、２個をＣＯＰＤとして誤って予測した。２４個のＣＯＰＤサンプルの間で、遺伝子シグネチャは、１６個をＣＯＰＤサンプルとして正しく予測したが、８個を対照として誤って予測した。

【００５６】

本発明の実装は、特定の例を参照して特定して示され、説明されているが、本開示の精神および範囲から逸脱することなく、形態および詳細における種々の変更がそれに行われ得ることが、当業者によって理解されるべきである。

【図１】

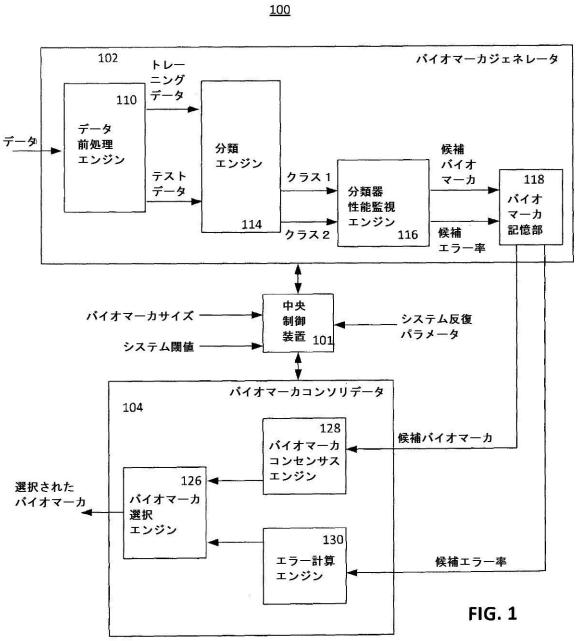

FIG. 1

【図２】

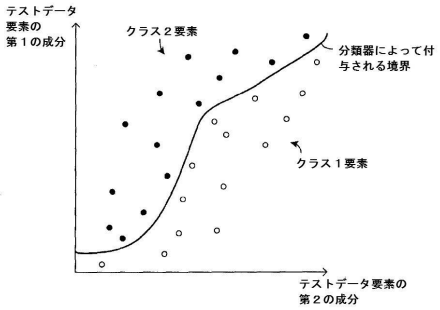

FIG. 2

【図３－１】

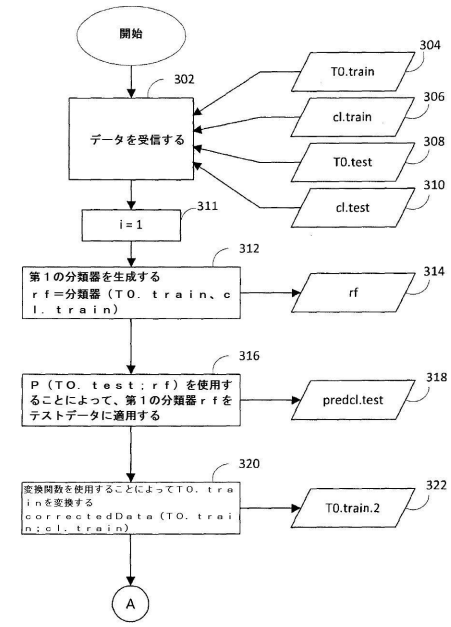

FIG. 3, SHEET 1

【図３－２】

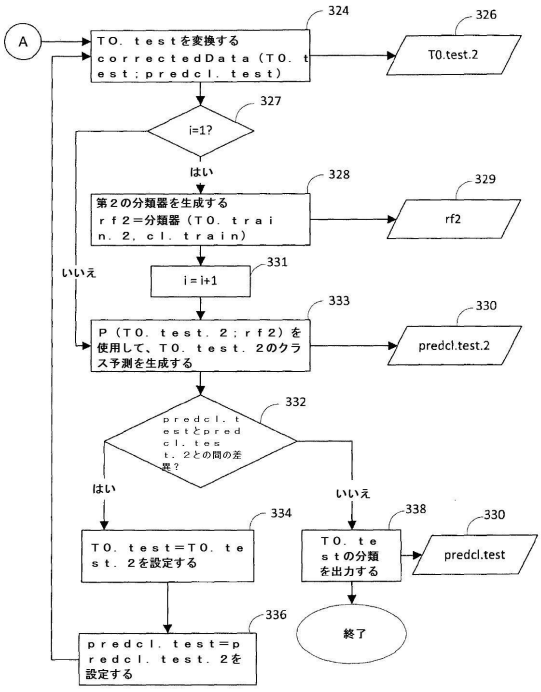

FIG. 3, SHEET 2

【図４】

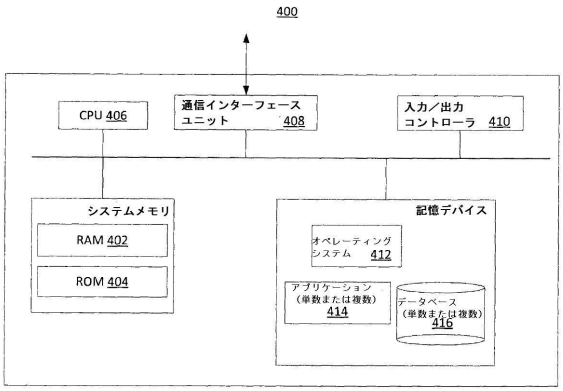

FIG. 4

【図５】

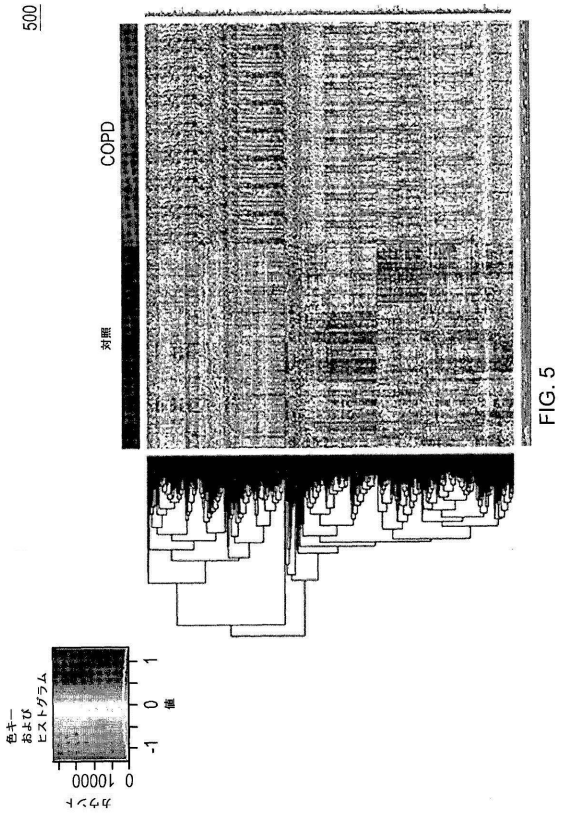

FIG. 5

フロントページの続き

(72)発明者　マルティン,　フロリアン
スイス国　ツェーハー－２０３４　プスー,　シュマン　ドゥ　ロレー　１
(72)発明者　シアン,　ヤン
スイス国　ツェーハー－２０００　ヌーシャテル,　リュ　ドゥ　ロシェ　２４

審査官　笠田　和宏

(56)参考文献　特表２０１２－５０１１８３（ＪＰ，Ａ）
特開２００９－０７５７３７（ＪＰ，Ａ）
米国特許出願公開第２００６／００７４８２６（ＵＳ，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
ＩＰＣ　Ｇ０６Ｆ　１５／１８
１７／３０
１９／００－１９／２８
Ｇ０６Ｎ　３／００－　３／１２
７／０８－９９／００
Ｇ０６Ｑ　１０／００－９９／００