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

特許第6313757号
(P6313757)

(45)発行日　平成30年4月18日(2018.4.18)　(24)登録日　平成30年3月30日(2018.3.30)

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

請求項の数 15　(全 29 頁)

(21)出願番号	特願2015-517784 (P2015-517784)
(86)(22)出願日	平成25年6月21日 (2013.6.21)
(65)公表番号	特表2015-527635 (P2015-527635A)
(43)公表日	平成27年9月17日 (2015.9.17)
(86)国際出願番号	PCT/EP2013/062982
(87)国際公開番号	WO2013/190085
(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	61/662,812
(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】統合デュアルアンサンブルおよび一般化シミュレーテッドアニーリング技法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）既知のラベルのセットを有するトレーニングデータセットを受信するステップと、

（ｂ）第１の機械学習技法を前記トレーニングデータセットに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であって、前記第１の機械学習技法は、分類方法の第１のセットを識別し、各分類方法は、前記トレーニングデータセットについて投票を行う、ステップと、

（ｃ）前記第１の分類器に従って、前記トレーニングデータセット中の要素を分類することにより、前記トレーニングデータセットについて、予測されるラベルの第１のセットを取得するステップと、

（ｄ）前記予測されるラベルの第１のセットおよび前記既知のラベルのセットから、第１の客観値を計算するステップと、

（ｅ）複数の反復の各々について、

（ｉ）第２の機械学習技法を前記トレーニングデータセットに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２の分類器を生成するステップであって、前記第２の機械学習技法は、分類方法の第２のセットを識別し、各分類方法は、前記トレーニングデータセットについて投票を行う、ステップと、

（ｉｉ）前記第２の分類器に従って、前記トレーニングデータセット中の要素を分類することにより、前記トレーニングデータセットについて、予測されるラベルの第２のセットを取得するステップと、

（ｉｉｉ）前記予測されるラベルの第２のセットおよび前記既知のラベルのセットから、第２の客観値を計算するステップと、

（ｉｖ）前記第１の客観値と前記第２の客観値とを比較することにより、前記第２の分類器が前記第１の分類器よりも性能が優れているかどうかを決定するステップと、

（ｖ）前記第２の分類器が前記第１の分類器よりも性能が優れている場合に、前記予測されるラベルの第１のセットを前記予測されるラベルの第２のセットと置換し、前記第１の客観値を前記第２の客観値と置換し、ステップ（ｉ）に戻るステップと、

（ｆ）所望の数の反復が達せられた場合に、前記予測されるラベルの第１のセットを出力するステップと

を含む、方法。

【請求項２】

前記トレーニングデータセットは、集約トレーニングデータセットからトレーニングデータサンプルのサブセットを選択することによって形成され、前記方法は、前記集約トレーニングデータセットをブートストラッピングすることにより、複数のさらなるトレーニングデータセットを生成するステップと、各さらなるトレーニングデータセットについて、ステップ（ａ）～ステップ（ｆ）を繰り返すステップとをさらに含む、請求項１に記載の方法。

【請求項３】

前記ブートストラッピングは、均衡のとれたサンプルを伴って、または、均衡のとれたサンプルを伴わずに行われる、請求項２に記載の方法。

【請求項４】

前記方法は、

前記出力された予測されるラベルの第１のセットをもたらした前記分類器を識別するステップと、

テストデータセット中のサンプルを選択するステップであって、前記テストデータセットは、前記トレーニングデータセットとは異なり、かつ、既知のラベルのセットを有さない、ステップと、

前記識別された分類器を使用することにより、前記選択されたサンプルについてラベルを予測するステップと

をさらに含む、請求項１～３のいずれかに記載の方法。

【請求項５】

前記分類方法の第１のセットは、分類方法の集約セットのサブセットを選択するように第１のランダムベクトルを使用することによって取得され、

前記第１のランダムベクトルは、分類方法の前記集約セットに対応する二進値のセットを含み、

各二進値は、前記集約セットにおける前記対応する分類方法が前記分類方法の第１のセットに含まれるかどうかを示し、

前記分類方法の第２のセットは、二進値の異なるセットを含む第２のランダムベクトルを使用することによって取得される、請求項１～４のいずれかに記載の方法。

【請求項６】

前記第２のランダムベクトルは、均衡のとれたブートストラッピングを行うべきかどうかを示すフラグ変数、ブートストラップの数、分類方法のリスト、遺伝子のリスト、または、それらの組み合わせをさらに含む、請求項５に記載の方法。

【請求項７】

前記第２の客観値は、前記予測されるラベルの第２のセットおよび前記既知のラベルのセットから査定されるマシューズ相関係数に対応する、請求項１～６のいずれかに記載の方法。

【請求項８】
前記第２の客観値を計算する前記ステップは、シミュレーテッドアニーリング方法を実装するステップを含む、請求項１～７のいずれかに記載の方法。
【請求項９】
前記第２の客観値を計算する前記ステップは、シミュレーテッドアニーリング方法を実装するステップを含み、前記シミュレーテッドアニーリング方法は、前記第２のランダムベクトルを取得するように前記第１のランダムベクトルの１つ以上の値を更新するステップを含む、請求項５に記載の方法。
【請求項１０】
前記第１のランダムベクトルの前記１つ以上の値を更新するステップは、前記第２のランダムベクトルを取得するように前記第１のランダムベクトルの各要素をランダムに更新するステップを含む、請求項９に記載の方法。
【請求項１１】
（１）前記第２の客観値が前記第１の客観値よりも小さい場合、および、（２）前記第２の客観値が前記第１の客観値よりも大きいときに、ランダム値が、前記第１の客観値と前記第２の客観値とから計算される確率値よりも小さい場合、前記第２の分類器が前記第１の分類器よりも性能が優れていることを決定するステップをさらに含む、請求項１～１０に記載の方法。
【請求項１２】
前記確率値は、制御パラメータｑ、前記第１の客観値、前記第２の客観値、および、冷却式から計算される温度値から計算される、請求項１１に記載の方法。
【請求項１３】
前記第２の分類器は、線形判別分析、サポートベクトルマシンベースの方法、ランダムフォレスト方法、および、ｋ最近傍方法を含む群から選択される、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、請求項１～１３のいずれかに記載の方法の１つ以上のステップを前記プロセッサに実行させる、コンピュータプログラム製品。
【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た処理デバイス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、前記処理デバイスに請求項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６１／６６２，８１２号（発明の名称「Ｓｙｓｔｅｍｓ ａｎｄ Ｍｅｔｈｏｄｓ ｆｏｒ Ｇｅｎｅｒａｔｉｎｇ Ｂｉｏｍａｒｋｅｒ Ｓｉｇｎａｔｕｒｅｓ ｗｉｔｈ Ｉｎｔｅｇｒａｔｅｄ Ｄｕａｌ Ｅｎｓｅｍｂｌｅ ａｎｄ Ｇｅｎｅｒａｌｉｚｅｄ Ｓｉｍｕｌａｔｅｄ Ａｎｎｅａｌｉｎｇ Ｔｅｃｈｎｉｑｕｅｓ」、２０１２年6月２１日出願）に対する３５ Ｕ．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】
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させ、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に、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】

本明細書において、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システム、コンピュータプログラム製品、および、方法が、説明される。特定すると、出願人らは、方法と遺伝子セットデータとの組み合わせが、個別方法のみよりもテストデータの良好な予測を提供でき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、要素を２つ以上のクラスに分類するための１つ以上のそのような技法を含む方法を実装する。特定すると、統合デュアルアンサンブル（ｉｎｔｅｇｒａｔｅｄ　ｄｕａｌ　ｅｎｓｅｍｂｌｅ）およびシミュレーテッドアニーリング技法を使用して、バイオマーカシグネチャが生成される。この技法は、データセットを再サンプリングし、デュアルアンサンブル方法を使用して表現型を予測することを伴う。特定すると、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、一組の分類方法およびデータサンプルを示すランダムベクトルを形成するステップを含む。ランダムベクトルは、反復して摂動させられ、異なる摂動に対応する異なる客観値（ｏｂｊｅｃｔｉｖｅ　ｖａｌｕｅ）が、計算される。

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を受信するステップと、

（ｂ）第１の機械学習技法を前記トレーニングデータセットに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、

（ｃ）前記第１の分類器に従って、前記トレーニングデータセット中の要素を分類することによって、第１のトレーニングクラスセットを生成するステップと、

（ｄ）前記トレーニングクラスセットに基づいて、第１の客観値を計算するステップと、

（ｅ）複数の反復の各々について、

（ｉ）第２の機械学習技法を前記トレーニングデータセットに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２の分類器を生成するステップと、

（ｉｉ）前記第２の分類器に従って、前記トレーニングデータセット中の要素を分類することによって、第２のトレーニングクラスセットを生成するステップと、

（ｉｉｉ）前記トレーニングクラスセットに基づいて、第２の客観値を計算するステップと、

（ｉｖ）前記第１の客観値と前記第２の客観値とを比較するステップと、

（ｖ）ステップ（ｉｖ）における比較に基づいて、前記第１のトレーニングクラスセットを前記第２のトレーニングクラスセットと置換し、前記第１の客観値を前記第２の客観値と置換し、ステップ（ｉ）に戻るステップと、

（ｆ）所望の数の反復が達せられた場合に、前記第１のトレーニングクラスセットを出力するステップと

を含む、方法。

（項目２）

前記方法は、複数のトレーニングデータセットについて前記ステップ（ａ）～（ｆ）を繰り返すステップをさらに含み、前記複数のトレーニングデータセットの中の各トレーニングデータセットは、集約トレーニングデータセットをブートストラップすることによって生成される、項目１に記載の方法。

（項目３）

前記ブートストラッピングは、均衡のとれたサンプルを伴って、または、均衡のとれたサンプルを伴わずに行われる、項目２に記載の方法。

（項目４）

テストデータセット中のサンプルを選択するステップと、前記出力された第１のトレーニングクラスセットに対応する前記分類器を使用することにより、前記選択されたサンプルと関連付けられる値を予測するステップとをさらに含む、項目１～３のいずれかに記載の

方法。
（項目５）
前記第２の分類器は、ランダムベクトルを適用することにより前記第２の分類器と関連付けられる分類スキームについてのパラメータを識別することによって生成され、前記ランダムベクトルは、少なくとも１つの二進値を含む、項目１～４のいずれかに記載の方法。
（項目６）
前記パラメータは、均衡のとれたブートストラッピングを行うべきかどうかを示すフラグ変数、ブートストラップの数、分類方法のリスト、遺伝子のリスト、または、それらの組み合わせを含む、項目５に記載の方法。
（項目７）
前記第２の客観値を計算する前記ステップは、マシューズ相関係数に基づく、項目１～６のいずれかに記載の方法。
（項目８）
前記第２の客観値を計算する前記ステップは、二進一般化シミュレーテッドアニーリング方法を実装するステップを含む、項目１～７のいずれかに記載の方法。
（項目９）
前記二進一般化シミュレーテッドアニーリング方法は、前記分類スキームについてのパラメータを識別するように、前記ランダムベクトルの１つ以上の値を局所的に摂動させるステップを含む、項目８に記載の方法。
（項目１０）
前記ランダムベクトルの前記１つ以上の値を局所的に摂動させるステップは、前記ランダムベクトルの各要素をランダムに更新することにより、更新されたランダムベクトルを取得するステップと、前記更新されたランダムベクトルを使用して、更新された第２の客観値を計算するステップと、確率値と乱数との間の比較に基づいて、前記更新された第２の客観値を受理するステップとを含む、項目９に記載の方法。
（項目１１）
前記ランダムベクトルの前記１つ以上の値を局所的に摂動させるステップは、各反復について前記ランダムベクトルの１つの要素を変更するステップを含む、項目９に記載の方法。
（項目１２）
前記第１のトレーニングクラスセットを前記第２のトレーニングクラスセットと置換し、前記第１の客観値を前記第２の客観値と置換する前記ステップは、冷却式に基づく、項目１～１１のいずれかに記載の方法。
（項目１３）
前記第２の分類器は、線形判別分析、サポートベクトルマシンベースの方法、ランダムフォレスト方法、および、ｋ最近傍方法を含む群から選択される、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、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前記プロセッサに実行させる、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た処理デバイス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、前記処理デバイスに項目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。
【０００９】
特定の局面において、本明細書で説明されるシステムおよび方法は、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ための手段および方法を含む

。本方法は、トレーニングデータセットを受信するステップを含み得る。トレーニングデータセットは、集約データセットを発見（トレーニング）セットと検証（テスト）セットとに分離することによって決定され得る。例えば、集約データセットは、複数のソースから一緒にプールされるデータを含んでもよく、集約データセットは、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され得る。本方法はさらに、第１の機械学習技法をトレーニングデータ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の分類器を生成するステップを含み得る。例えば、機械学習技法は、サポートベクトルマシン（ＳＶＭ）、または、特徴選択のための任意の好適な技法に対応し得る。第１のトレーニングクラスセットが、第１の分類器に従ってトレーニングデータセット中の要素を分類することによって生成される。特定すると、第１の分類器は、データセット中の各サンプルを生理学的状態（例えば、罹患または疾患なし等）に割り当てる分類規則に対応し得る。第１の分類器は、ＳＶＮ、ネットワークベースのＳＶＭ、ニューラルネットワークベースの分類器、ロジスティック回帰分類器、決定木ベースの分類器、線形判別分析技法、ランダムフォレスト分析技法、任意の他の好適な分類方法、または、前述のものの組み合わせを使用する分類器等の複数の分類方法を組み合わせ得る。

【００１０】

第１の客観値が、トレーニングクラスセットに基づいて計算される。特定すると、客観値を計算するために、二進一般化シミュレーテッドアニーリング方法（ｂｉｎａｒｙ ｇｅｎｅｒａｌｉｚｅｄ ｓｉｍｕｌａｔｅｄ ａｎｎｅａｌｉｎｇ ｍｅｔｈｏｄ）が、使用され得る。ランダムベクトルは、その要素として、使用されるべき分類技法を定義する一組のパラメータを含み得る。ランダムベクトルによって定義される本技法は、第１の客観値を計算するために使用される。次いで、複数の反復について、第２の機械学習技法が、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ように、トレーニングデータセットに適用され、第２のトレーニングクラスセットが、第２の分類器に従ってトレーニングデータセット中の要素を分類することによって生成される。特定すると、第２の分類器は、第１の分類器を定義するために使用されるランダムベクトルをランダムに摂動させ、かつ、第２の分類器を定義するためにランダムベクトルのランダム摂動を使用することによって、生成され得る。さらに、第２のトレーニングクラスセットに基づく第２の客観値が計算され、第１の客観値と第２の客観値とが比較される。第１の客観値と第２の客観値との間の比較に基づいて、第１のトレーニングクラスセットは、第２のトレーニングクラスセットと置換され得、第１の客観値は、第２の客観値によって置換され得、次の反復が開始される。反復は、所望の数の反復が達せられ、かつ、第１のトレーニングクラスセットが出力されるまで繰り返される。

【００１１】

上記で説明される方法の特定の実施形態において、本方法のステップは、複数のトレーニングデータセットについて繰り返され、複数のトレーニングデータセットの中の各トレーニングデータセットは、集約トレーニングデータセットをブートストラップすることによって生成される。ブートストラッピングは、均衡のとれたサンプルを伴って、または、均衡のとれたサンプルを伴わずに行われ得る。均衡のとれたサンプルを伴って、または、均衡のとれたサンプルを伴わずにブートストラップするかどうかは、ランダムベクトルが摂動させられるときに値が更新され得るランダムベクトルの中の二進要素によって決定され得る。置換を伴って、または、置換もしくはいくつかのブートストラップを伴わずに、サンプルの集約セットからサンプルのサブセットをサンプリングするかどうか等の他のブートストラップパラメータが、要素としてランダムベクトルに含まれ得る。本方法の特定の実施形態において、サンプルが、テストデータセットの中で選択され、出力された第１のトレーニングクラスセットに対応する分類器は、選択されたサンプルと関連付けられる値を予測するために使用される。方法の特定の実施形態において、第２の分類器は、第２の分類器と関連付けられる分類スキームについてのパラメータを識別するように、ランダムベクトルを適用することによって生成され、そのランダムベクトルは、少なくとも１つの二進値を含む。本方法の特定の実施形態において、ランダムベクトルのパラメータは、

均衡のとれたブートストラッピングを行うべきかどうかを示すフラグ変数、ブートストラップの数、分類方法のリスト、遺伝子のリスト、または、それらの組み合わせを含む。

【００１２】

本方法の特定の実施形態において、第２の客観値を計算するステップは、マシューズ相関係数に基づく。特定すると、客観値は、１と、結果のマシューズ相関係数との間の差に対応し得る。マシューズ相関係数は、複合性能スコアとして使用され得る性能測定基準である。本方法の特定の実施形態において、第２の客観値を計算するステップは、二進一般化シミュレーテッドアニーリング方法を実装するステップを含む。本方法の特定の実施形態において、二進一般化シミュレーテッドアニーリング方法は、分類スキームについてのパラメータを識別するように、ランダムベクトルの１つ以上の値を局所的に摂動させるステップを含む。本方法の特定の実施形態において、ランダムベクトルの１つ以上の値を局所的に摂動させるステップは、更新されたランダムベクトルを取得するように、ランダムベクトルの各要素をランダムに更新するステップと、更新されたランダムベクトルを使用して、更新された第２の客観値を計算するステップと、確率値と乱数との間の比較に基づいて、更新された第２の客観値を受理するステップとを含む。本方法の特定の実施形態において、ランダムベクトルの１つ以上の値を局所的に摂動させるステップは、各反復についてランダムベクトルの１つの要素を変更するステップを含む。

【００１３】

本方法の特定の実施形態において、第１のトレーニングクラスセットを第２のトレーニングクラスセットと置換し、第１の客観値を第２の客観値と置換するステップは、冷却式に基づく。特定すると、ランダムベクトルに対して大幅な摂動を行うことによって、二進一般化シミュレーテッドアニーリング方法において客観値を減少させることが、望ましくあり得る。シミュレーテッドアニーリングにおいて、冷却をシミュレートするように、人工温度値が徐々に低減される。１つの点（すなわち、ランダムベクトルについての第１の組の値）から別の点（すなわち、ランダムベクトルについての第２の組の値）までの試験ジャンプ距離（ｔｒｉａｌ　ｊｕｍｐ　ｄｉｓｔａｎｃｅ）をシミュレートするために、訪問分布（ｖｉｓｉｔｉｎｇ　ｄｉｓｔｒｉｂｕｔｉｏｎ）が、シミュレーテッドアニーリングにおいて使用される。試験ジャンプは、第２の客観値が第１の客観値よりも小さいかどうか、および、受理確率に基づいて受理される。二進一般化シミュレーテッドアニーリング方法は、客観値を最小限化するためのグローバルミニマムを識別するために使用される。本方法の特定の実施形態において、第２の分類器は、線形判別分析、サポートベクトルマシンベースの方法、ランダムフォレスト方法、および、ｋ最近傍方法を含む群から選択される。

【００１４】

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】

【００１５】

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。

【００１６】

【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、データサンプルの分類および分類規則の決定を描写するグラフである。

【図３】図３は、デュアルアンサンブル方法の流れ図である。

【図４】図４は、データセットを構築するための方法の流れ図である。

【図５】図５は、結果ベクトルおよび客観値を生成するための方法の流れ図である。

【図６】図６は、二進一般化シミュレーテッドアニーリング方法を初期化するための方法の流れ図である。

【図７】図７は、二進一般化シミュレーテッドアニーリング方法において客観値を減少させるための方法の流れ図である。

【図８】図８は、二進一般化シミュレーテッドアニーリング方法において客観値をさらに減少させるための方法の流れ図である。

【図９】図９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１０】図１０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】

【００１７】

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、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。

【００１８】

本明細書で説明されるシステムおよび方法は、統合デュアルアンサンブル（ｉｎｔｅｇｒａｔｅｄ　ｄｕａｌ　ｅｎｓｅｍｂｌｅ）およびシミュレーテッドアニーリング技法を用いてバイオマーカシグネチャを生成するための技法を含む。本技法は、データセットを再サンプリングし、デュアルアンサンブル方法を使用して表現型を予測することを伴う。特定すると、本明細書で説明されるシステムおよび方法は、一組の分類方法およびデータサンプルを示すランダムベクトルを形成することと、そのランダムベクトルを反復して摂動させることと、異なる摂動に対応する異なる客観値を計算することとを含む。

【００１９】
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。

【００２０】
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を２つ以上のクラス（例えば、罹患データおよび非罹患、感染しやすい、および、免疫がある等）のうちの１つに分類するための分類器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。

【００２１】
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、特に、分類器１１４の動作は、図２～８への参照によってさらに詳細に説明される。

【００２２】
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。

【００２３】

データ前処理エンジン１１０は、１つ以上のデータセットを受信する。概して、データは、サンプル中の複数の異なる遺伝子の発現値、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、あるいは、ｍｉＲＮＡ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。そのような修飾ポリペプチドの各々のレベルは、別々に決定され、データセットにおいて表され得る。

【００２４】

分類器１１４は、データ前処理エンジン１１０から１つ以上のデータのセットを受信する。特定の実施形態において、分類器１１４は、データを分類するように分類規則を生成する。図２は、そのような分類規則２００を図式的に描写する。分類器１１４は、データセットを２つのクラスのうちのいずれか１つに割り当てるように、分類規則を適用し得る。例えば、分類器１１４は、データセットを疾患または対照のいずれかに割り当てるように、分類を適用し得る。

【００２５】

特定の実施形態において、図３～８に関連して説明されるように、分類器１１４は、分類規則を生成するために、一般化シミュレーテッドアニーリング方法と組み合わせられたデュアルアンサンブル技法を使用する。特定すると、分類器１１４は、サポートベクトルマシン（ＳＶＭ）、ネットワークベースのＳＶＭ、ニューラルネットワークベースの分類器、ロジスティック回帰分類器、決定木ベースの分類器、線形判別分析技法および／またはランダムフォレスト分析技法を用いる分類器、または、任意の他の好適な分類方法等の

複数の分類方法を組み合わせ得る。アンサンブル分類方策は、最適な分類を識別するために、複数の多様な分類方法にわたって投票プロセスを使用し得る。複数の分類方法を組み込むことによって、アンサンブル技法は、少量のデータセットに過剰適合する可能性を低減する。このようにして、他の技法と比較して、アンサンブル技法を使用することによって、少量のデータセットが、より効率的に使用され得る。さらに、複数の分類方法のアンサンブルを使用して、特に、アンサンブルの中の複数の分類方法が相互に異なる場合に、単一の分類方法を使用することと比較して、強化された分類を可能にする。

【００２６】

加えて、データ前処理エンジン１１０から受信されるデータは、より良好な分類精度を提供しながら、全体的な多様性をさらに増加させるように摂動させられ得る。データの摂動の例は、図４、図７、および、図８に関連してさらに詳細に説明される。

【００２７】

本明細書で説明されるように、分類器１１４は、分類規則を生成するために、アンサンブル技法および一般化シミュレーティングアニーリング方法を使用し、生物情報学における適用に関連して説明される。しかしながら、本明細書で説明されるシステムおよび方法は、概して、特徴選択または抽出等の任意の大規模計算技法に適用され得る。

【００２８】

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、分類器１１４の性能を分析し得る。特定すると、分類器１１４の性能を分析する場合、分類器性能監視エンジン１１６は、１つ以上の候補バイオマーカのロバスト性または性能を分析していてもよい。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。候補バイオマーカおよび対応する性能測定基準は、バイオマーカ記憶部１１８に記憶され得る。

【００２９】

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。

【００３０】

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。

【００３１】

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）。

【００３２】

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。

【００３３】

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。

【００３４】

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象から、血液サンプルが収集され、それによって、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。

【００３５】

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。

【００３６】

分類器１１４は、データ前処理エンジン１１０から１つ以上の候補バイオマーカおよび１つ以上のデータのセットを受信し得る。分類器１１４は、データセットを２つのクラスのうちのいずれか１つに割り当てるように、分類規則を適用し得る。例えば、分類器１１４は、データセットを疾患または対照のいずれかに割り当てるように、分類を適用し得る。特定の実施形態において、分類器１１４は、サポートベクトルマシン（ＳＶＭ）分類器、ネットワークベースのＳＶＭ、ニューラルネットワークベースの分類器、ロジスティック回帰分類器、決定木ベースの分類器、線形判別分析技法および／またはランダムフォレスト分析技法を用いる分類器を含み得る。分類器１１４およびそれぞれのエンジンの動作は、図２～８への参照によってさらに詳細に説明される。

【００３７】
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、分類器１１４の性能を分析し得る。特定すると、分類器１１４の性能を分析する場合、分類器性能監視エンジン１１６は、１つ以上の候補バイオマーカのロバスト性または性能を分析していてもよい。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。候補バイオマーカおよび対応する性能測定基準は、バイオマーカ記憶部１１８に記憶され得る。
【００３８】
前述のように、ＣＣＵ　１０１はまた、バイオマーカジェネレータ１０２において生成されて記憶された候補バイオマーカに基づいて、好適かつ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の動作を制御し得る。バイオマーカコンソリデータ１０４は、バイオマーカ記憶部１１８から１つ以上の候補バイオマーカを受信するバイオマーカコンセンサスエンジン１２８を含む。バイオマーカコンセンサスエンジン１２８は、新しいバイオマーカシグネチャについて、１つ以上の候補バイオマーカ内で頻繁に発生する遺伝子を選択し得る。新しいバイオマーカシグネチャは、Ｎが、バイオマーカの所望のサイズ、バイオマーカの最大許容サイズ、バイオマーカの最小許容サイズ、または、最大サイズと最小サイズとの間のサイズであるＮ個の遺伝子を含み得る。特定の実施形態において、数Ｎは、ユーザ選択可能であり得、かつ、所望に応じて調整可能であり得る。
【００３９】
図３は、投票方法を使用して表現型クラスを予測するために分類器１１４によって使用される方法３００の流れ図である。示されるように、方法３００は、Ｋ個のデータセットを構築するステップ（ステップ３０２）と、Ｍ個の分類方法を識別するステップ（ステップ３０６）と、Ｋ個のデータセットの各々の中でＧ個のサンプルを識別するステップ（ステップ３１２）とを含む。方法３００はさらに、Ｋ個のデータセット、Ｍ個の分類方法、および、Ｇ個のサンプルにわたって反復を行うステップを含む３つの反復ループを含み、Ｇは、テストデータセットのサンプルサイズである。特定すると、各反復において、分類方法ｊが、表現型を予測するようにデータセットｉ中のサンプルｌに適用され（ステップ３１８）、ｉ＝１，２，．．．Ｋ、ｊ＝１，２，．．．Ｍ、かつ、ｌ＝１，２，．．．Ｇである。
【００４０】
ステップ３０２において、分類器１１４は、Ｋ個のデータセットを構築する。分類器は、Ｋ個のデータセットを構築するために、図４に描写される方法を使用し得る。特定すると、分類器１１４は、完全なデータセットの複数のデータセットを形成するためにブートストラッピング集約方法（ｂｏｏｔ　ｓｔｒａｐｐｉｎｇ　ａｇｇｒｅｇａｔｉｏｎ　ｍｅｔｈｏｄ）を使用し得る。ステップ３０４において、データセットに適用されるラベルを表すデータセット反復パラメータｉが、１に初期化される。
【００４１】
ステップ３０６において、分類器１１４は、Ｍ個の分類方法を識別する。分類器１１４は、外部ソースから分類方法を受信し得るか、または、分類方法が、いくつかの入力に基づいて分類器１１４によって生成され得る。例として、分類器１１４は、方法３０８のリストに基づいて、Ｍ個の分類方法を識別し得る。方法の例は、線形判別分析、サポートベクトルマシンベースの方法、ランダムフォレスト方法（Ｂｒｅｉｍａｎ，　Ｍａｃｈｉｎｅ　Ｌｅａｒｎｉｎｇ，　４５（１）：５－３２　（２００１））、ＰＡＭＲ（Ｔｉｂｓｈｉｒａｎｉ　ｅｔ　ａｌ．，　Ｐｒｏｃ　Ｎａｔｌ　Ａｃａｄ　Ｓｃｉ　ＵＳＡ，　９９（１０）：６５６７－６５７２　（２００２））、または、ｋ最近傍方法（Ｂｉｓｈｏｐ，　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋｓ　ｆｏｒ　Ｐａｔｔｅｒｎ　Ｒｅｃｏｇｎｉｔｉｏｎ，　ｅｄ．　Ｏ．Ｕ．　Ｐｒｅｓｓ，　１９９５）を含む。任意の数の分類方法が、使用され、考慮され得る。ステップ３１０において、分類方法に適用されるラベルを表

す方法反復パラメータｊが、１に初期化される。ステップ３１６において、データサンプルに適用されるラベルを表すサンプル反復パラメータｌが、１に初期化される。各データサンプルは、個人、遺伝子、または、任意の他の好適なデータ点を表し得る。

【００４２】

ステップ３１２において、分類器１１４は、テストデータセット中のｌ番目のサンプルを選択し、ステップ３１８において、分類器１１４は、分類器を構築するように分類方法ｊをデータセットｉに適用し、テストデータ中のサンプルｌを予測する。サンプルｌの予測は、表現型の予測に対応し得る。いくつかの実施形態において、表現型は、フラグ変数（すなわち、個人が表現型を発現すると予測される場合は１、そうでなければ０）であり得る。しかしながら、概して、表現型は、任意の数の値をとり得る。特定すると、表現型予測は、値として３次元行列Ｐ（ｉ，ｊ，ｌ）３２０に記憶され得る。

【００４３】

決定ブロック３２２において、分類器１１４は、最後のデータセットが考慮されているかどうか、または、同等に、ｉ＝Ｋであるかどうかを決定する。ｉがＫよりも小さい場合、分類器１１４は、ステップ３２４でデータセット反復パラメータｉをインクリメントし、ステップ３１８に戻って新しいデータセットについての表現型を予測する。

【００４４】

Ｋ個全てのデータセットが考慮された後、分類器１１４は、決定ブロック３２６へ進んで、最後の分類方法が適用されているかどうか、または、同等に、ｊ＝Ｍであるかどうかを決定する。ｊがＭよりも小さい場合、分類器１１４は、ステップ３２８で方法反復パラメータｊをインクリメントし、ステップ３１８に戻って新しい分類方法についての表現型を予測する。

【００４５】

Ｋ個全てのデータセットが考慮され、Ｍ個全ての分類方法が適用された後、分類器１１４は、現在のデータサンプルｌについてのＫ×Ｍ個の表現型予測を有する。これらの表現型予測は、投票と考えられ得、任意の種類の投票計数方法が、一組のＫ×Ｍ個の表現型予測を表す複合投票に到達するために使用され得る。

【００４６】

決定ブロック３３２において、分類器は、Ｇ個全てのデータサンプルが考慮されているかどうか、または、同等に、ｌ＝Ｇであるかどうかを決定する。

【００４７】

図４は、データセットを構築するための方法４００の流れ図であり、図３におけるステップ３０２で分類器１１４によって使用され得る。概して、方法４００は、より大きいデータセットの各サブセットである複数のデータセットを生成するための方法を提供する。データサブセットは、大きいデータセット中のサンプルのサブセットをランダムに選択することを伴うブートストラップ集約（「バギング」）方法によって形成され得る。サンプルのサブセットは、置換を伴うかまたは伴わずに、選択され得る。示されるように、方法４００は、データを受信するステップ（ステップ４４０）と、置換を伴わずにブートストラッピングを行うことが望ましいかどうかを決定するステップ（決定ブロック４５４）とを含む。そうである場合、Ｗ個のサンプルが、データセットを形成するように各クラスからランダムに選択され得る（ステップ４５６）。代替として、Ｈ個のサンプルが、データセットを形成するようにトレーニングデータから置換を伴ってランダムに選択され得る（ステップ４６０および４６６）。Ｈの値は、トレーニングデータセットのサンプルサイズに対応し得る。上記のステップは、図３に関連して説明される各データセットｉが考慮されるまで繰り返される。

【００４８】

ステップ４４０において、分類器１１４は、データを受信する。データは、２つのクラス（すなわち、クラス１サンプル４４２およびクラス２サンプル４４４）にソートされるサンプル、ブートストラップパラメータ４４６、および、結果として生じるデータセットｉ（すなわち、データサブセット）のサイズとクラス（すなわち、クラス１またはクラス

２）のサイズとの間の比ｓ　４４８を含み得る。例として、ブートストラップパラメータ４４６は、置換を伴うかまたは伴わずにブートストラップするかどうかを示す変数、および、ブートストラップデータセットの数（すなわち、Ｋ）を含み得る。データ４４２、４４４、４４６、および、４４８は、Ｋ個のデータセットを構築するために分類器１１４によって使用され得る。

【００４９】

ステップ４５２において、データセット反復パラメータｉが、１に初期化される。反復パラメータｉは、データセットに適用されるラベルを表す。

【００５０】

決定ブロック４５４において、分類器１１４は、均衡のとれたサンプルを用いてブートストラップすることが望ましいかどうかを決定する。特定すると、分類器１１４は、均衡のとれたサンプルを用いたブートストラッピングが望ましいかどうかを決定するように、ブートストラップパラメータ４４６等の変数を使用し得る。概して、均衡のとれたサンプルを用いたブートストラッピングは、Ｋ個全てのデータセットにわたって各サンプル点の発生の総数が同一であることを確実にする。

【００５１】

均衡のとれたブートストラッピングが望ましい場合、分類器１１４は、ステップ４５０へ進んでデータセットサイズWを決定する。特定すると、例えば、Ｗ＝最小値｛サイズ（クラス１サンプル），サイズ（クラス２サンプル）｝＊ｓのように、サイズWは、比ｓ　４４８に依存し得る。特定すると、比ｓは、０から１の間の値であり得る。ステップ４５６において、トレーニングデータセットからのW個のサンプルが、均衡のとれたサンプルとともにランダムに選択され、データセットｉ　４５８を形成する。反復パラメータｉが１よりも大きい場合、ステップ４５６におけるW個のサンプルの選択は、ブートストラッピングが均衡を保たれるように、以前に形成されたデータセットに依存し得る。

【００５２】

代替として、均衡のとれたサンプルを用いたブートストラッピングが望ましくない場合、分類器１１４は、ステップ４６０へ進んで、置換を伴ってトレーニングデータセットからＨ個のサンプルをランダムに選択する。選択されたサンプルは、データセットｉ　４６４を形成する。

【００５３】

図４に描写されるように、均衡のとれたブートストラッピングが、サイズWを有するデータセットをもたらす一方で、均衡のとれたサンプルを伴わずにデータをブートストラップすることは、サイズＨを有するデータセットをもたらす。しかしながら、概して、サイズWを有するデータセットについての均衡のとれたサンプルを伴わないブートストラッピング、または、サイズＨを有するデータセットについての均衡のとれたブートストラッピング等の、方法の任意の好適な組み合わせが使用され得る。加えて、置換方法を伴わないブートストラッピングもまた使用され得る。

【００５４】

現在のデータセットｉが形成された後、分類器１１４は、決定ブロック４７０へ進んで、最後のデータセットが形成されているかどうか、または、同等に、ｉ＝Ｋであるかどうかを決定する。そうでない場合、ステップ４７２において、データセット反復パラメータｉがインクリメントさせられ、分類器１１４は、決定ブロック４５４へ進んで次のデータセットを形成し始める。

【００５５】

図５は、結果ベクトルおよび客観値を生成するための方法の流れ図である。概して、方法５００は、ランダムベクトルＸに対応する客観値を計算する方法を提供する。方法５００で描写されるように、ランダムベクトルＸは、二進ベクトル（ｂｉｎａｒｙ　ｖｅｃｔｏｒ）Ｘであり、置換を伴ってブートストラップするかどうかに関する情報（５０６）、ブートストラップの数（５１０）、分類方法のリスト（５１４）、および、データサンプルのリスト（５１８）を含む。これらのデータに基づいて、予測行列が形成され（ステッ

プ５２０）、主要クラスが決定される（ステップ５２４）。分類器１１４は、全てのデータサンプルが考慮されるまで、データサンプルにわたって反復を行い、客観値が、データサンプルについての決定された主要クラスに基づいて計算される（ステップ５３２）。

【００５６】

ステップ５０２において、分類器１１４は、二進ランダムベクトルＸを受信する。例において、ベクトルＸは、二進値のリストであり得る。二進値は、均衡のとれたブートストラッピングを行うかどうか、ブートストラップの数（すなわち、Ｋ）、分類方法のリスト、および／または、遺伝子のリストを示し得る。特定すると、ブートストラップの数は、ゼロ値またはゼロではない値（すなわち、例えば６０）のいずれかをとり得る。この場合、ブートストラップの数に対応するベクトルＸの中の二進値は、ブートストラップの数がゼロであるか、または、ゼロではないかを示し得る。乱数値ジェネレータ、または、乱数値を生成するための任意の他の好適な方法によって、乱数値が、生成され得る。本明細書で説明されるように、ランダムベクトルＸは、ベクトルの中の各値が２つの値のうちの１つ（すなわち、０または１）であることを意味する二進ベクトルである。しかしながら、概して、ランダムベクトルＸの中の値は、任意の数の値のうちの１つにあり得る。分類器１１４は、ベクトルＸの中の乱数値に基づいて、種々のパラメータを識別する。例として、分類器１１４は、ステップ５０４において均衡のとれたサンプルを用いてサンプリングするかどうかを示すフラグ５０６についての値、ステップ５０８でブートストラップの数５１０、ステップ５１２において分類方法のリスト５１４、および、ステップ５１６において遺伝子のリスト５１８を識別する。

【００５７】

識別された種々のパラメータに基づいて、ステップ５２０で、分類器１１４は、予測行列を生成する。

【００５８】

ステップ５２２において、データサンプルに適用されるラベルを表すサンプル反復パラメータｌが、１に初期化される。

【００５９】

ステップ５２４において、分類器１１４は、主要クラスＰ（．，．，ｌ）を決定する。特定すると、分類器１１４は、Ｋ×Ｍ個の表現型予測を識別するように、方法３００におけるステップ３０２～３３０を通してパース（ｐａｒｓｅ）を行い、主要クラスＰ（．，．，ｌ）を決定するように、Ｋ×Ｍ個の予測について多数決を行ってもよい。概して、一組のＫ×Ｍ個の予測に基づいて複合予測を生成するための任意の他の好適な方法が、主要クラスを決定するように使用され得る。主要クラスは、入力として結果ベクトル５２６に記憶され得る。

【００６０】

決定ブロック５２８において、分類器１１４は、サンプル反復パラメータｌがデータサンプルの総数Ｇに等しいかどうかを決定する。そうでない場合、反復パラメータｌがステップ５３０でインクリメントさせられ、主要クラスが、次のデータサンプルについて決定される。

【００６１】

主要クラスが一組のＧ個のサンプルの中の各サンプルについて決定された後、分類器１１４は、ステップ５３２へ進んで客観値を計算する。客観値は、結果ベクトル５２６の中の、結果として生じた一組の入力に基づいて計算され得る。特定すると、複合性能スコアが、性能測定基準の平均であり得る。方法５００で描写されるように、客観値５３２は、１と結果のマシューズ相関係数（ＭＣＣ）との間の差として計算される。ＭＣＣは、複合性能スコアとして使用され得る性能測定基準である。特定すると、ＭＣＣは、－１と＋１との間の値であり、本質的に、観察された二進分類と予測された二進分類との間の相関係数である。ＭＣＣは、下記の式を使用して計算され得る。

【数１】

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

式中、ＴＰ：真陽性、ＦＰ：偽陽性、ＴＮ：真陰性、ＦＮ：偽陰性である。しかしながら、概して、一組の性能測定基準に基づいて複合性能測定基準を生成するための任意の好適な技法が、客観値を計算するために使用され得る。

【００６２】

図６～８は、二進一般化シミュレーテッド方法のステップを通してパースを行うための方法の流れ図である。概して、二進一般化シミュレーテッドアニーリング方法は、図５で説明されるような客観値についての最適値（すなわち、グローバルミニマム）を識別するために使用され得る。本明細書で説明されるように、二進一般化シミュレーテッドアニーリング方法が、図３で説明されるデュアルアンサンブル方法と併せて使用される。特定すると、図５で説明されるようなランダムベクトルＸが、最適な客観値を識別するように種々の方法で摂動させられる。図６は、二進一般化シミュレーテッドアニーリング方法を初期化するための流れ図である。図７は、客観値を減少させるようにランダムベクトルＸの種々の成分をランダムに摂動させるための流れ図である。図８は、客観値をさらに減少させるようにランダムベクトルＸを局所的に摂動させるための流れ図である。換言すると、図７で描写される方法が、ランダムベクトルＸの大幅な摂動を生成する一方で、図８で描写される方法は、ランダムベクトルＸの軽微な摂動を生成する。

【００６３】

図６は、二進一般化シミュレーテッドアニーリング方法を初期化するための方法６００の流れ図である。方法６００は、いくつかのパラメータを初期化し、ランダム二進ベクトルＸ（１）を生成する。特定すると、ステップ６４０、６４２、６４４において、分類器１１４は、それぞれ、パラメータｔ、ｙ、および、１へのカウントを初期化する。パラメータｔは、図７および８に関連して説明されるように、時間間隔に対応し、好適な客観値が決定されるときにインクリメントさせられる。反復パラメータｙは、行われるべき大幅な摂動の数に対応し、図７に関連してさらに詳細に説明される。パラメータカウントは、現在のベクトルＸの摂動バージョンが生成されているかどうかを記録するためのパラメータに対応し、図７に関連してさらに詳細に説明される。ステップ６４６において、分類器１１４は、ランダム二進ベクトルＸを生成する。

【００６４】

ステップ６４８において、パラメータＤが設定される。パラメータＤは、摂動させられるように選択されるＸの中の成分の数に対応する。特定すると、ステップ６４８において、パラメータＤは、０．２＊Ｃに設定され、Ｃは、二進ベクトルＸの長さに対応する。

【００６５】

ステップ６５０において、分類器１１４は、結果ベクトルおよび客観値を生成する。特定すると、分類器１１４は、結果ベクトル５２６および客観値５３４を生成するために、図５で描写される方法を使用し得る。しかしながら、概して、複合性能測定基準を表す客観値を決定するための任意の好適な方法が、使用され得る。客観値を生成した後、分類器１１４は、図７のステップへ進んで、ランダムベクトルＸを摂動させることによって客観値を減少させる。

【００６６】

図７は、ベクトルＸに大幅な摂動を行うことによって、二進一般化シミュレーテッドアニーリング方法において客観値を減少させるための方法の流れ図である。シミュレーティングアニーリング方法において、人工温度が導入され（Ｔ（ｔ＝１））、冷却をシミュレートするように徐々に低減される。１つの点から第２の点まで（すなわち、１つのランダムベクトルＸ（１）から別のランダムベクトルＸ（２）まで）の試験ジャンプ距離をシミュレートするために、訪問分布が、シミュレーテッドアニーリングにおいて使用される。試験ジャンプは、ランダムベクトルＸ（２）に対応する、結果として生じる客観値が、ラ

ンダムベクトルＸ（１）に対応する客観値よりも小さいかどうか、および、下記で定義されるような受理確率に基づいて受理される。本明細書で説明されるように、グローバルミニマムの場所を特定するために（すなわち、客観値を最小限化するために）、二進一般化シミュレーテッドアニーリング方法が、使用される。しかしながら、概して、最急降下、共役勾配、シンプレックス、および、モンテカルロ法等の任意の好適なアルゴリズムが、使用され得る。

【００６７】

図６で描写される方法を使用してシミュレーションを初期化した後、分類器１１４は、ステップ７６０において、ベクトルＸ（１）のＤ個の成分を選択し始める。ベクトルＸ（１）のＤ個の成分は、ランダムに選択され得るか、または、ベクトルＸ（１）のＤ個の成分を選択する任意の他の好適な方法が、行われ得る。ステップ７６２において、カウント変数が２に設定される。ステップ７６４において、変更されたＤ個の成分を有する元のランダムベクトルＸ（１）に対応する第２のランダム二進ベクトルＸ（２）が、生成される。

【００６８】

ステップ７６６において、分類器１１４は、第２のベクトルＸ（２）についての結果ベクトル７６８および客観値７７０を生成する。特定すると、分類器１１４は、結果ベクトルおよび客観値を生成するために、図５で描写される方法を使用し得る。しかしながら、概して、複合性能測定基準を表す客観値を決定するための任意の好適な方法が、使用され得る。

【００６９】

第２の結果ベクトルおよび第２の客観値を生成した後、分類器は、決定ブロック７７２において、カウント変数が２に等しいことを決定し、決定ブロック７７６へ進んで、第１の客観値（すなわち、ランダムベクトルＸ（１）に対応する）と第２の客観値（すなわち、ランダムベクトルＸ（２）に対応する）とを比較する。

【００７０】

第２の客観値が第１の客観値よりも小さくない場合、これは、第１のベクトルＸ（１）が、第２のベクトルＸ（２）としてより良好であるかまたは平しい相関をもたらしたことを意味する。この場合、分類器は、ステップ７７８へ進んで確率Ｐを計算する。特定すると、確率Ｐは、第２の客観値を受理する確率に対応し、下記の方程式に基づく。

【数２】

$$P=\min\{1,[1-(1-q_a)\beta\partial E]^{\frac{1}{1-q_a}}\}$$

式中、 $\partial E=obj(2)-obj(1)$

$$\beta=\frac{1}{T_{q_v}(t)}$$

q_aは、確率Ｐを受理するための制御パラメータである。

T_{q_v}は、温度値である。

【００７１】

本明細書で説明されるように、確率Ｐは、一般化シミュレーテッドアニーリング方法において使用される確率に対応するが、概して、任意の好適な確率値が、使用され得る。ステップ７８６において、０以上１以下の乱数ｒが生成される。乱数ｒは、一様分布、または、任意の他の好適な分布から生成され得、ｒは、決定ブロック７８８において確率Ｐと比較される。

【００７２】

Ｐがｒ以上である場合、これは、第２の客観値が第１の客観値よりも小さくなかったとしても、第２の客観値を受理する確率が高いことを意味する。この場合、分類器１１４は、ステップ７９０、７９２へ進んで、第１のベクトルＸ（１）として第２のベクトルＸ（２）を記憶し、第１の客観値として第２の客観値を記憶する。

【００７３】
代替として、決定ブロック７７６において、分類器１１４が、第２の客観値が第１の客観値よりも小さいことを決定する場合、これは、ベクトルＸ（２）が、より良好な相関、または、より良好な性能をもたらしたことを意味する。したがって、分類器は、ステップ７９０へ直接進んで、ベクトルＸ（２）でベクトルＸ（１）を更新し、ステップ７９２へ進んで、第２の客観値で第１の客観値を更新する。ステップ７９４において、分類器１１４は、カウント変数を１に等しく設定する。
【００７４】
代替として、決定ブロック７８８において、分類器１１４が、ｒがＰよりも大きいことを決定する場合、これは、第２の客観値を受理する確率が低いことを意味し、それによって、ステップ７９０、７９２が迂回され、ベクトルＸ（１）および第１の客観値が対応する第２の値によって上書きされない。この場合、分類器１１４は、ステップ７９４へ進み、カウント変数を１に等しく設定する。
【００７５】
カウント変数１を１に再設定した後、分類器１１４は、反復パラメータｙが値Ｌと比較される決定ブロック７９６へ進む。値Ｌは、軽微な摂動を行うように図８で描写される方法へ進む前に行われるべき大幅な摂動の最大数に対応する。反復パラメータｙがＬに等しくない場合、分類器１１４は、決定ブロック７７２およびステップ７７４へ進んで反復パラメータｙをインクリメントし、ステップ７６０～７６４においてベクトルＸの大幅な摂動を行う。上記で説明されるステップは、所望の数の大幅な摂動Ｌが行われるまで繰り返される。図７で描写されるように、行われるべき大幅な摂動の数は、固定数Ｌである。しかしながら、Ｌについての値は、任意の数の要因に依存し得る。例えば、分類器１１４は、客観値の収束に基づいて、大幅な摂動の総数に達したことを決定し得る。別の例において、決定ブロック７７６において、いかなる第２の客観値も固定数の最近の比較で第１の客観値よりも小さいことが見出されなかった場合に、大幅な摂動の総数に達し得る。概して、大幅な摂動が行われること、および、分類器１１４が図８へ進んで軽微な摂動を行い得ることを決定するように、任意の好適な方法が、使用され得る。
【００７６】
図８は、ベクトルＸに軽微な摂動を行うことによって、二進一般化シミュレーテッドアニーリング方法において客観値をさらに減少させるための方法の流れ図である。特定すると、方法８００は、ステップ８０２から始まり、ベクトルＸ（１）の長さに等しい変数Ｃを設定する。ステップ８０４において、分類器１１４は、反復パラメータｃを１に初期化し、改善フラグ変数（ｉｍｐｒｏｖｅ　ｆｌａｇ　ｖａｒｉａｂｌｅ）を偽に設定する。
【００７７】
ステップ８０６において、分類器１１４は、X_{temp}を生成するようにＸ（１）のｃ番目のビットを反転させることによって、ベクトルＸ（１）に軽微な摂動を行う。特定すると、Ｘ（１）は、長さＣの二進ベクトルであり、X_{temp}は、ｃ番目のビットを除いてＸ（１）とほぼ同一である。
【００７８】
ステップ８０８において、分類器１１４は、一時ベクトル（ｔｅｍｐｏｒａｒｙ　ｖｅｃｔｏｒ）X_{temp}に対する結果ベクトル８１０および客観値８１２を生成する。特定すると、分類器１１４は、一時結果ベクトルおよび一時客観値を生成するために、図５で描写される方法を使用し得る。しかしながら、概して、複合性能測定基準を表す客観値を決定するための任意の好適な方法が、使用され得る。
【００７９】
決定ブロック８１４において、第１の客観値は、一時客観値と比較される。一時客観値が第１の客観値よりも小さい場合、これは、摂動バージョンX_{temp}が元のベクトルＸ（１）よりも良好な性能をもたらしたことを意味する。この場合、分類器１１４は、ステップ８１６へ進んで、摂動バージョンX_{temp}でベクトルＸ（１）を上書きし、ステップ８１８へ進んで、一時客観値で第１の客観値を上書きし、ステップ８１９へ進んで、改

善フラグ変数を真に設定する。

【００８０】

決定ブロック８２０において、分類器１１４は、ベクトルＸ（１）の中の各ビットが少なくとも１回（すなわち、ステップ８０６において）反転させられているかどうか、または、同等に、反復パラメータｃがＸ（１）　Ｃのサイズに等しいかどうかを決定する。そうでない場合、分類器１１４は、ステップ８２２へ進んで反復パラメータｃをインクリメントさせ、ステップ８０６へ進んでｃ番目のビットを反転させる。

【００８１】

そうでなければ、分類器１１４が、決定ブロック８２０において、反復パラメータｃがベクトルＸ（１）　Ｃの長さに等しいことを決定する場合、分類器１１４は、決定ブロック８２２へ進んで、さらなる改善が所望されるかどうかを決定する。特定すると、分類器１１４は、さらなるビットフリッピングが望ましいかどうかを決定するように、改善フラグ変数の値を識別し得る。例えば、改善フラグ変数が真である場合、分類器１１４は、ステップ８０４へ戻って反復パラメータｃを１に再初期化し、改善フラグ変数を偽に再初期化する。

【００８２】

図８の描写された方法は、軽微な摂動（すなわち、ビットフリッピング）を行うプロセスが完了しているときを決定するように、改善フラグ変数を使用する。しかしながら、概して、任意の他の好適な方法もまた軽微な摂動が完了しているときを決定するように使用され得る。例えば、分類器１１４は、客観値がいくつかの閾値を下回ること、または、客観値と一時客観値との間の差がいくつかの閾値を下回ることを要求し得る。これらの要求が満たされない場合、分類器１１４は、ステップ８０６に戻って、別の一時客観値を生成するようにベクトルＸ（１）の別のビットを反転させてもよい。

【００８３】

分類器１１４が、最小客観値が識別されたことを決定した後、分類器１１４は、ステップ８２４、８２６へ進んで、それぞれにおいて、パラメータｔをインクリメントし、パラメータＤを減少させる。

【００８４】

ステップ８２８において、分類器１１４は、一般化シミュレーテッドアニーリングにおいて一般的に使用される冷却式によって、温度Ｔを計算する。しかしながら、任意の好適な式が使用され得る。

【数３】

$$T_{q_v}(t) \leftarrow T_{q_v}(1)\frac{2^{q_v-1}-1}{(1+t)^{q_v-1}-1}$$

式中、q_vは、分布関数の曲率を定義するパラメータである。

【００８５】

決定ブロック８３０において、分類器１１４は、$T_{q_v}(t)$がT_Lよりも小さいかどうかを決定する。T_Lについての値は、閾値を表し、$T_{q_v}(t)$についての値がT_Lを下回る場合、方法８００が終了し、現在のランダムベクトルＸ（１）が、最適な分類として使用される。

【００８６】

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。

【００８７】

図９は、図２～８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９００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９００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９００にわたって実装され得る。

【００８８】

コンピューティングデバイス９００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９１０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含む。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　９０２）と、少なくとも１つの読み取り専用メモリ（ＲＯＭ　９０４）とを含む。これらの要素は全て、中央処理ユニット（ＣＰＵ　９０６）と通信し、コンピューティングデバイス９００の動作を促進する。コンピューティングデバイス９００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９００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９００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９００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９において、コンピューティングデバイス９００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。

【００８９】

コンピューティングデバイス９００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９０８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。

【００９０】

ＣＰＵ　９０６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　９０６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　９０６は、通信インターフェースユニット１００８および入力／出力コントローラ９１０と通信し、それらを通して、ＣＰＵ　９０６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９０８および入力／出力コントローラ９１０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。

【００９１】

ＣＰＵ　９０６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　９０２、ＲＯＭ　９０４、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　９０６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　９０６は、通信インターフェースユニット９０８を介して、データ記憶デバイスに接続され得る。ＣＰＵ　９０６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。

【００９２】

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９００のためのオペレーティングシステム１０１２、（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　９０６に関して詳細に説明されるプロセスに従って、ＣＰＵ　９０６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９１４（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）９１６を記憶し得る。いくつかの局面において、データベース（単数または複数）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を含む。

【００９３】

オペレーティングシステム９１２およびアプリケーション９１４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　９０４またはＲＡＭ　９０２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　９０６に、本明細書に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。

【００９４】

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分類方法を行うことに関連する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９１２、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９１０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。

【００９５】

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９００のプロセッサ（または、本明細書に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。

【００９６】

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　９０６（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９００（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】

【００９７】

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).

【００９８】

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。均衡のと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１０に示す。図１０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１０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。遺伝子シグネチャは、１６個の対照サンプルのうちの１５個を正しく予測し、２４個のＣＯＰＤサンプルのうちの１４個を正しく予測した。

【００９９】

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、形態および詳細の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】

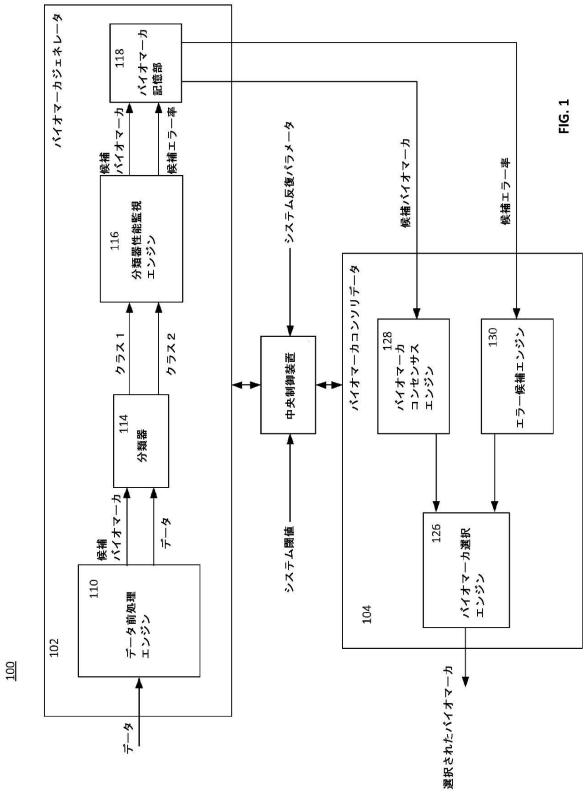

FIG. 1

【図２】

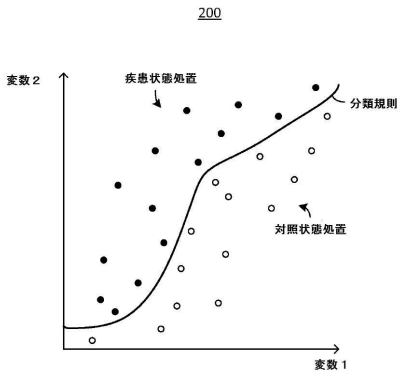

FIG. 2

【図３】

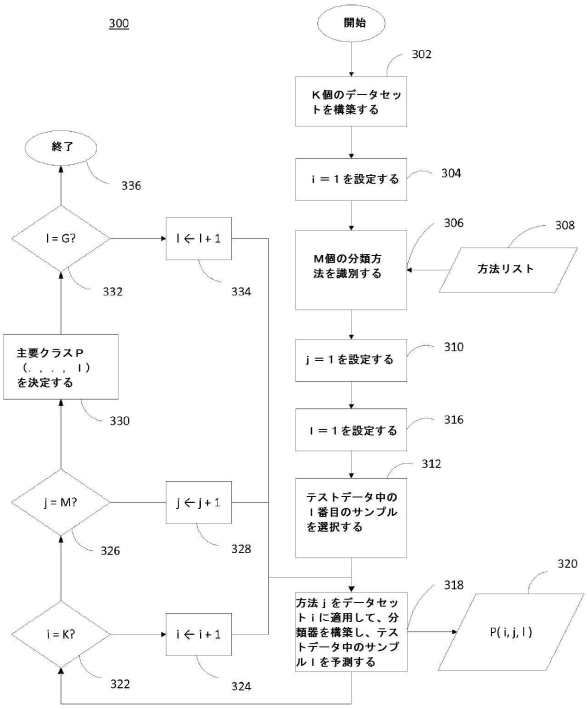

FIG. 3

【図４】

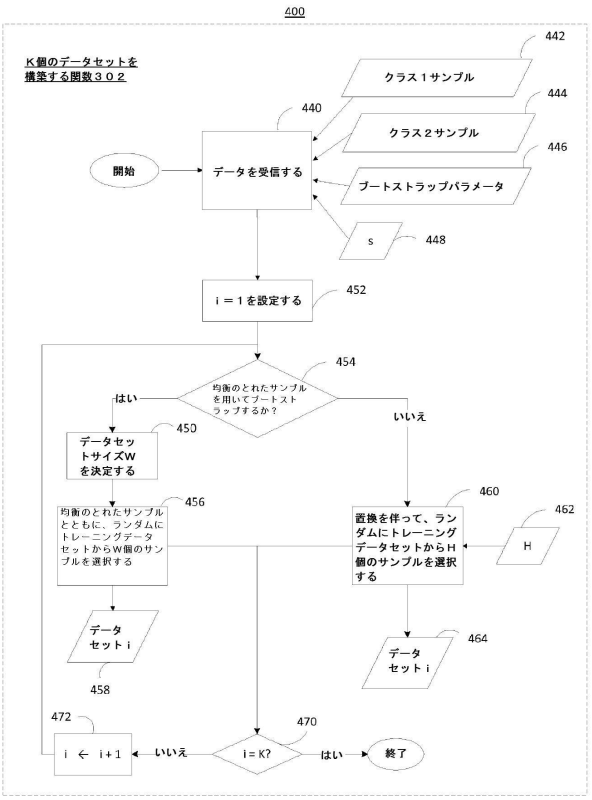

FIG. 4

【図５】

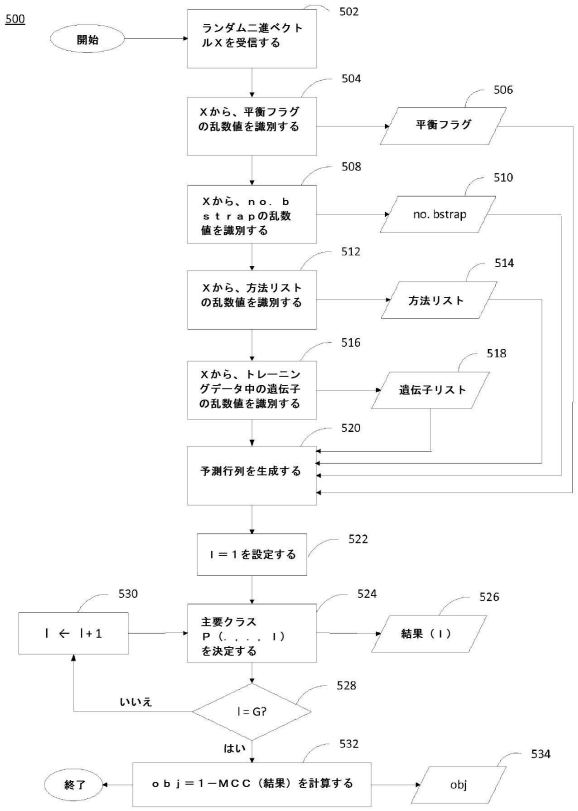

FIG. 5

【図６】

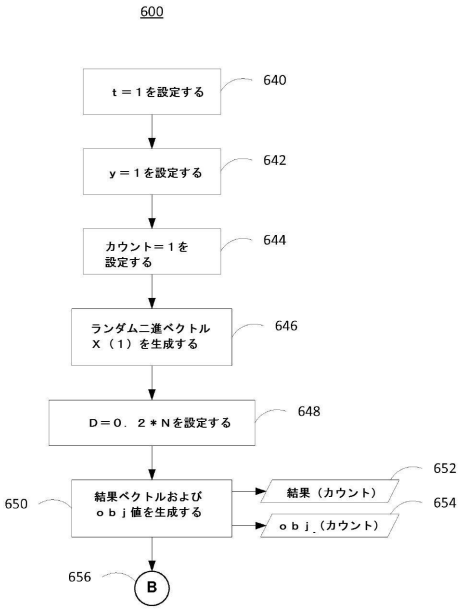

FIG. 6

【図７】

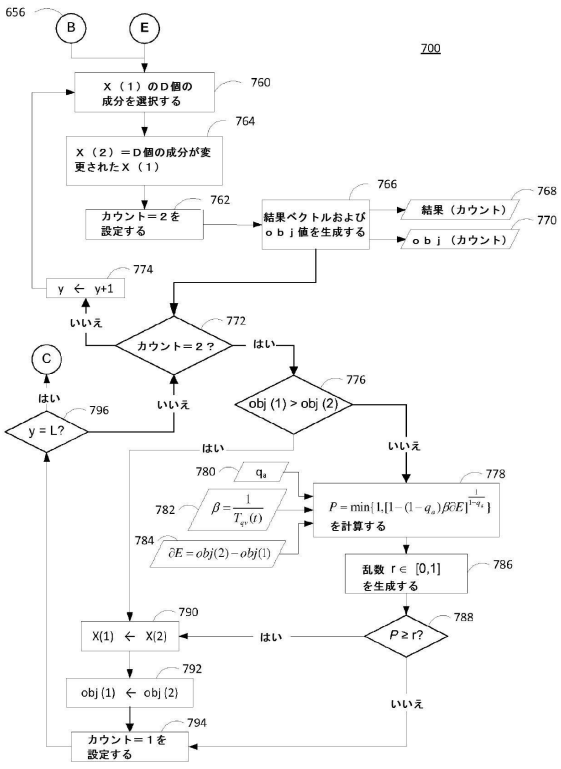

FIG. 7

【図８】

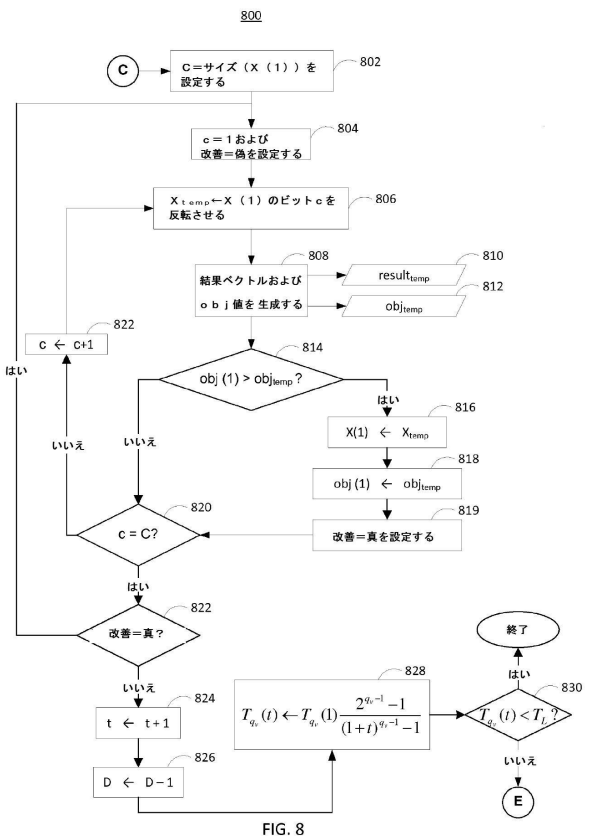

FIG. 8

【図９】

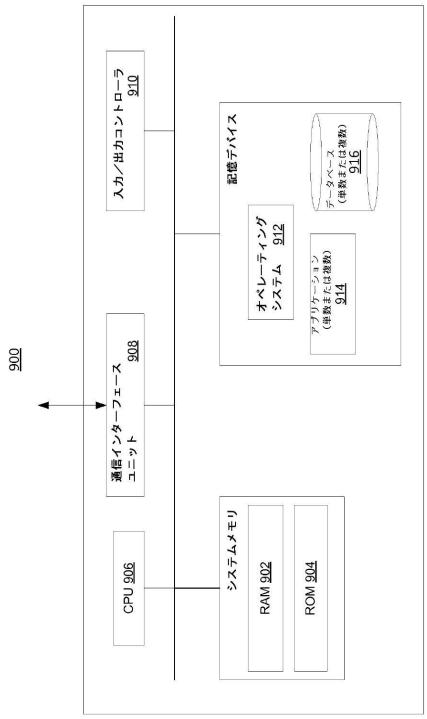

FIG. 9

【図１０】

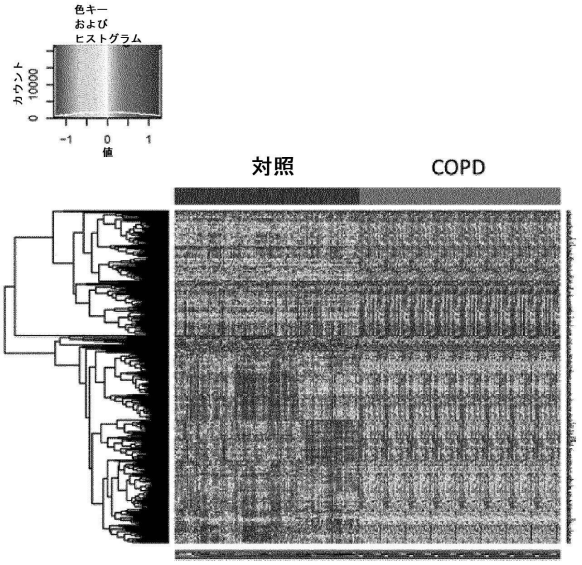

FIG. 10

フロントページの続き

(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
(72)発明者 ヘンク, ユリア
スイス国 ツェーハー－２０３５ コルセル, グラン－リュ ３５
(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

審査官 山内 裕史

(56)参考文献 特開２００９－２８２６８６（ＪＰ，Ａ）
特表２００５－５３８４３７（ＪＰ，Ａ）
特開２００８－０９０８３３（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １９／１０－１９／２８
Ｇ０６Ｆ １７／３０
Ｇ０６Ｎ ９９／００