

(19) 日本国特許庁(JP)

(12) 公表特許公報(A)

(11) 特許出願公表番号

特表2015-527635
(P2015-527635A)

(43) 公表日 平成27年9月17日(2015.9.17)

(51) Int. Cl.	F I	テーマコード (参考)
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

審査請求 未請求 予備審査請求 未請求 (全 44 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)
 (85) 翻訳文提出日 平成27年2月12日 (2015. 2. 12)
 (86) 国際出願番号 PCT/EP2013/062982
 (87) 国際公開番号 W02013/190085
 (87) 国際公開日 平成25年12月27日 (2013. 12. 27)
 (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)

(71) 出願人 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
 (71) 出願人 514323279
 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 ロシエ 24
 (71) 出願人 514322881
 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

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ハイオマールカシグネチャを生成するためのシステムおよび方法

(57) 【要約】

本明細書において、アンサンブル分類技法を使用してデータセットを分類するためのシステムおよび方法が説明されている。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、分類器が反復して生成され、そ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、トレーニングクラスセットが生成される。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、異なる分類器と関連付けられる客観値が、所望の数の反復が達せられ、かつ、最終トレーニングクラスセットが出力されるまで、比較される。

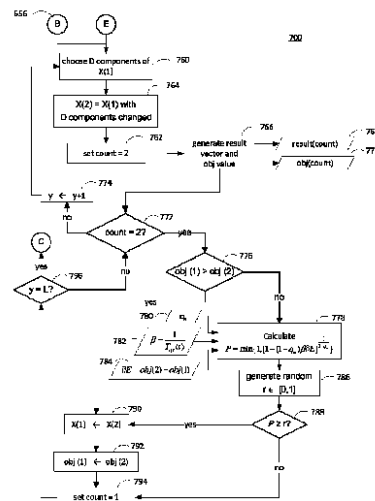


FIG. 7

【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

10

、

- (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、

- (i i) 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、

- (i i i) 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、

- (i v) 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(i v)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、

20

- (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。

【請求項 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、請求項1に記載の方法。

30

【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】

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、請求項1～3のいずれかに記載の方法。

【請求項 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、請求項1～4のいずれかに記載の方法。

40

【請求項 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、請求項5に記載の方法。

【請求項 7】

前記第2の客観値を計算する前記ステップは、マッシュアップ相関係数に基づく、請求項1～6のいずれかに記載の方法。

50

【請求項 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、請求項 1 ~ 7 のいずれかに記載の方法。

【請求項 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、請求項 8 に記載の方法。

【請求項 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、請求項 9 に記載の方法。

10

【請求項 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、請求項 9 に記載の方法。

【請求項 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、請求項 1 ~ 11 のいずれかに記載の方法。

20

【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。

30

【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。

40

【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なく

50

、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な可能性を提供する。特定の場において、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。

10

【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。

20

【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）。

30

【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、L I M M A、乳癌に関するバイオマーカを識別するためのF D A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（S V M）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。

40

【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

50

されることができ、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 (i n t e g r a t e d d u a l e n s e m b l e) およびシミュレーテッドアニ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 (o b j e c t i v e v a l u e) が、計算される。

10

20

30

40

50

【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 (S V M)、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、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、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リング方法 (b i n a r y g e n e r a l i z e d s i m u l a t e d a n n e a l i n g m e t h o d) 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。

10

20

30

40

50

【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。

10

20

30

40

50

【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一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一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 (i n t e g r a t e d d u a l e n s e m b l e) 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ー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。

10

【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、一つ以上のバイオマーカー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、バイオマーカージェネレータ102と、バイオマーカーコンソリデータ104とを含む。システム100はさらに、バイオマーカージェネレータ102およびバイオマーカーコンソリデータ104の動作の特定の局面を制御するための中央制御装置 (C C U) 101を含む。動作中に、遺伝子発現データ等のデータが、バイオマーカージェネレータ102で受信される。バイオマーカージェネレータ102は、複数の候補バイオマーカー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。バイオマーカーコンソリデータ104は、これらの候補バイオマーカー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ーを選択する。

20

【0020】

バイオマーカージェネレータ102は、データを処理して一組の候補バイオマーカー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、バイオマーカージェネレータ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。バイオマーカージェネレータ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) のうちの一つに分類するための分類器114を含む。バイオマーカージェネレータ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ーを識別し、一つ以上の候補バイオマーカーについて、候補エラー率を含み得る性能尺度を生成する。バイオマーカージェネレータ102はさらに、一つ以上の候補バイオマーカーおよび候補性能尺度を記憶するためのバイオマーカー記憶部118を含む。

30

40

【0021】

バイオマーカージェネレータは、自動的に制御またはユーザ操作され得る C C U 101 によって制御され得る。特定の実施形態において、バイオマーカージェネレータ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ー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ーを生成するために、バイオマーカージェネレータ102の動作は、複数回、反復され得る。C C U 101は、所望の数の候補バイオマーカーを含む一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、バイオマーカージェネレータ102の動作が反復され得

50

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エーション（*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

10

20

30

40

50

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

10

20

30

40

50

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一つに各々に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、一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。

10

【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。ヘテロジェニシティ（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、一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、一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）。

20

【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、インピッ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。

30

【0034】

一つの例において、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。

40

【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

50

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。

10

【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。

20

【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。

30

【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)、 $i = 1, 2, \dots, K$ 、 $j = 1, 2, \dots, M$ 、かつ、 $l = 1, 2, \dots, G$ である。

40

【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

50

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 (bootstrapping 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 $K \times M$ 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 $K \times M$ 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。

10

【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。

20

【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。

30

【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、 $W = \text{最小値}\{\text{サイズ(クラス 1 サンプル)}, \text{サイズ(クラス 2 サンプル)}\} * s$ 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。

40

【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

50

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つの値のうちの一つ (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一つにあり得る。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 $P(\cdot, \cdot, l)$ を決定する。特定すると、分類器 114 は、 $K \times M$ 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 $P(\cdot, \cdot, l)$ を決定するように、 $K \times M$ 個の予測について多数決を行ってもよい。概して、一組の $K \times M$ 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。

10

20

30

40

50

【 0 0 6 0 】

決定ブロック 5 2 8 において、分類器 1 1 4 は、サンプル反復パラメータ 1 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 1 がステップ 5 3 0 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。

【 0 0 6 1 】

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 1 1 4 は、ステップ 5 3 2 へ進んで客観値を計算する。客観値は、結果ベクトル 5 2 6 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 5 0 0 で描写されるように、客観値 5 3 2 は、1 と結果のマッシュアップ相関係数 (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最適な技法が、客観値を計算するために使用され得る。

【 0 0 6 2 】

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メーリング方法は、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メーリング方法が、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メ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。

【 0 0 6 3 】

図 6 は、二進一般化シミュレーテッドアニメーリング方法を初期化するための方法 6 0 0 の流れ図である。方法 6 0 0 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 6 4 0、6 4 2、6 4 4 において、分類器 1 1 4 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、最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 6 4 6 において、分類器 1 1 4 は、ランダム二進ベクトル X を生成する。

【 0 0 6 4 】

ステップ 6 4 8 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 6 4 8 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。

【 0 0 6 5 】

10

20

30

40

50

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。

10

20

【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。

30

【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。

40

【数 2】

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

式中、 $\partial E = \text{obj}(2) - \text{obj}(1)$

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

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_{qv} は、温度値である。

【0071】

10

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。

20

【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

30

【0075】

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。

40

【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

50

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 $X(1)$ 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 $X(1)$ 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プ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、ビットフリップング) を行うプロセスが完了しているときに決定するように、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に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を、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

10

20

30

40

50

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。

10

【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 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。

20

【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。特定の实装において、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。

30

【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。

40

【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る、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

50

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がなく、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

10

20

30

40

50

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。

10

【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（電氣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。

20

【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氣的、電磁的、または、光学的な信号として受信され得る。

30

40

【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).

10

【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.llda、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。マッシュ相関係数(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

20

30

40

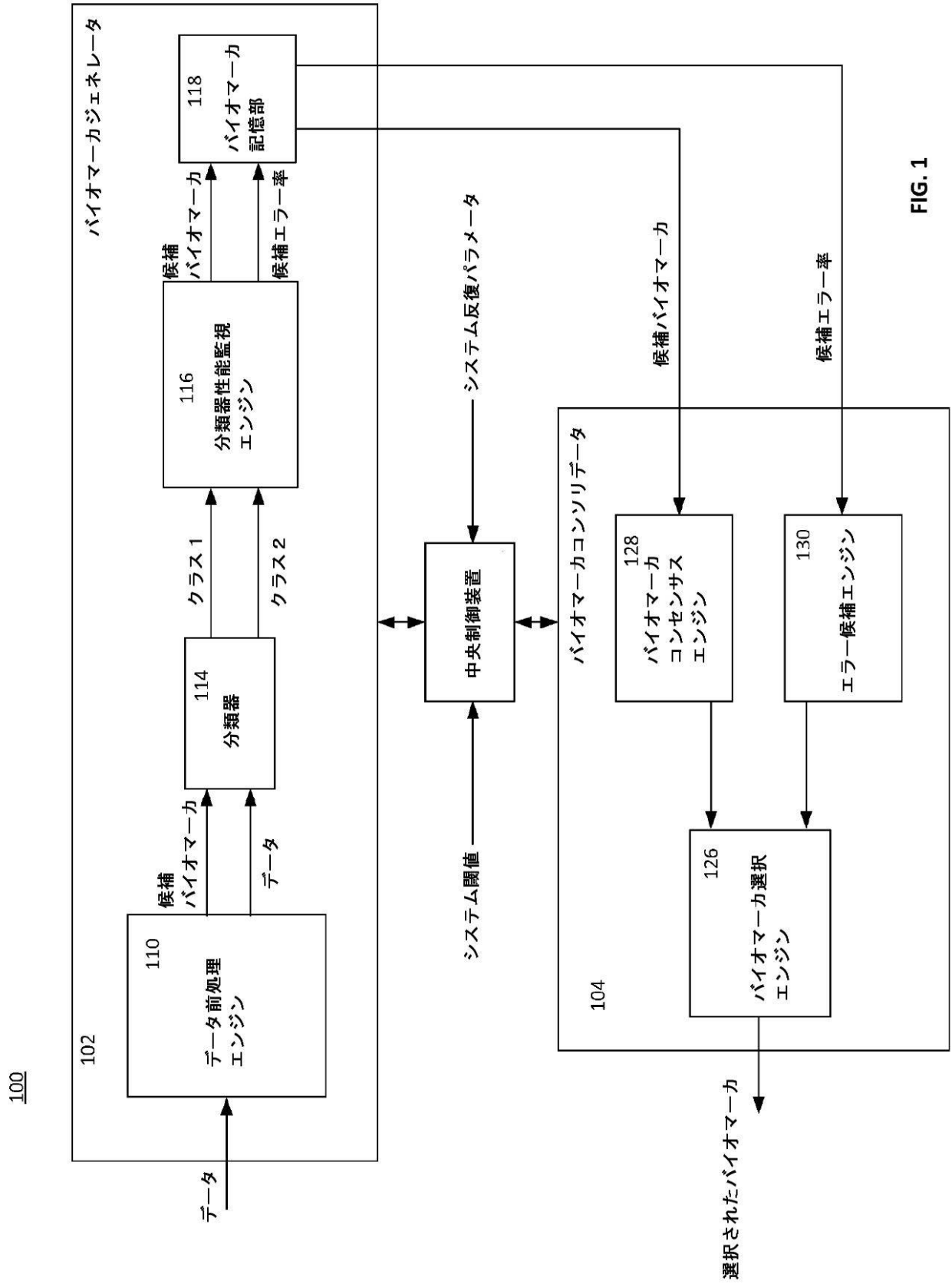
50

測した。

【 0 0 9 9 】

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 】



【 图 2 】

200

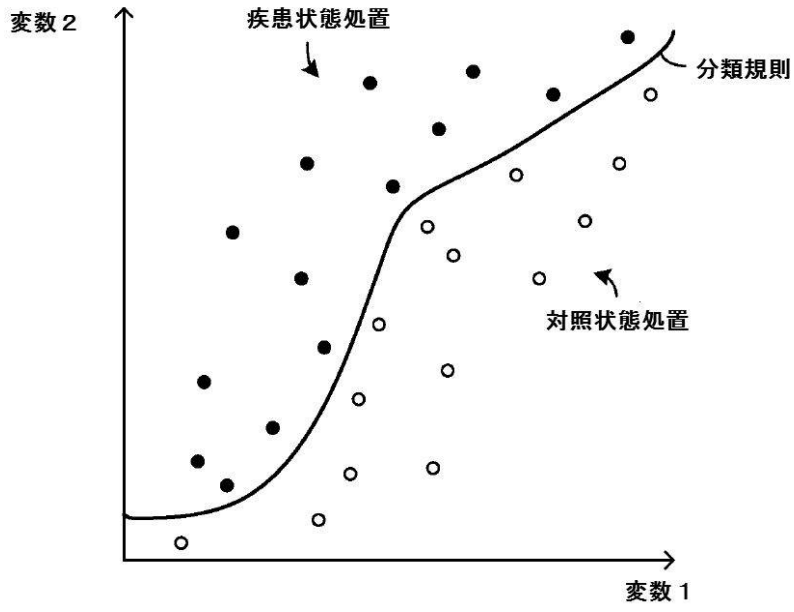


FIG. 2

【 図 3 】

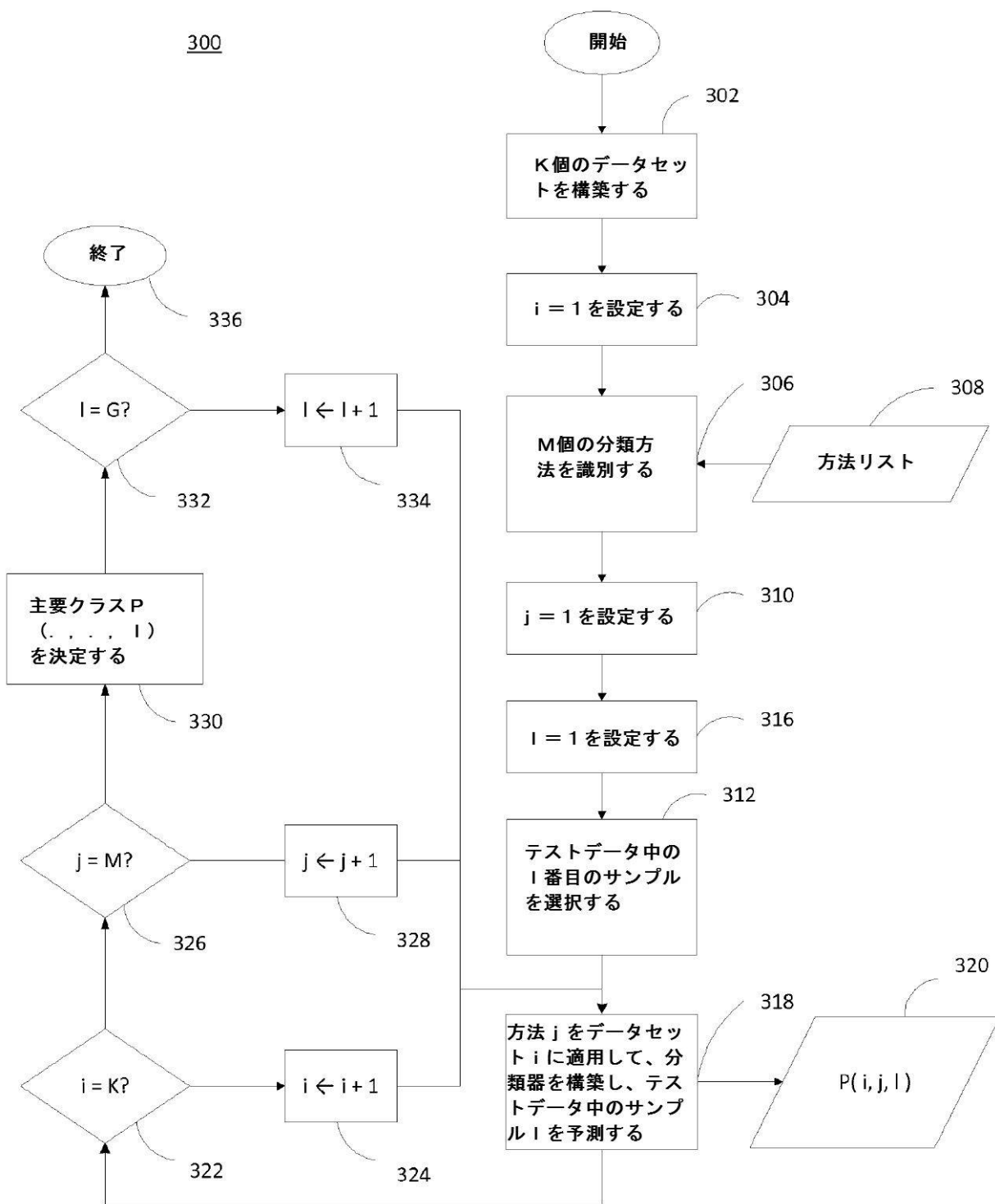


FIG. 3

【図4】

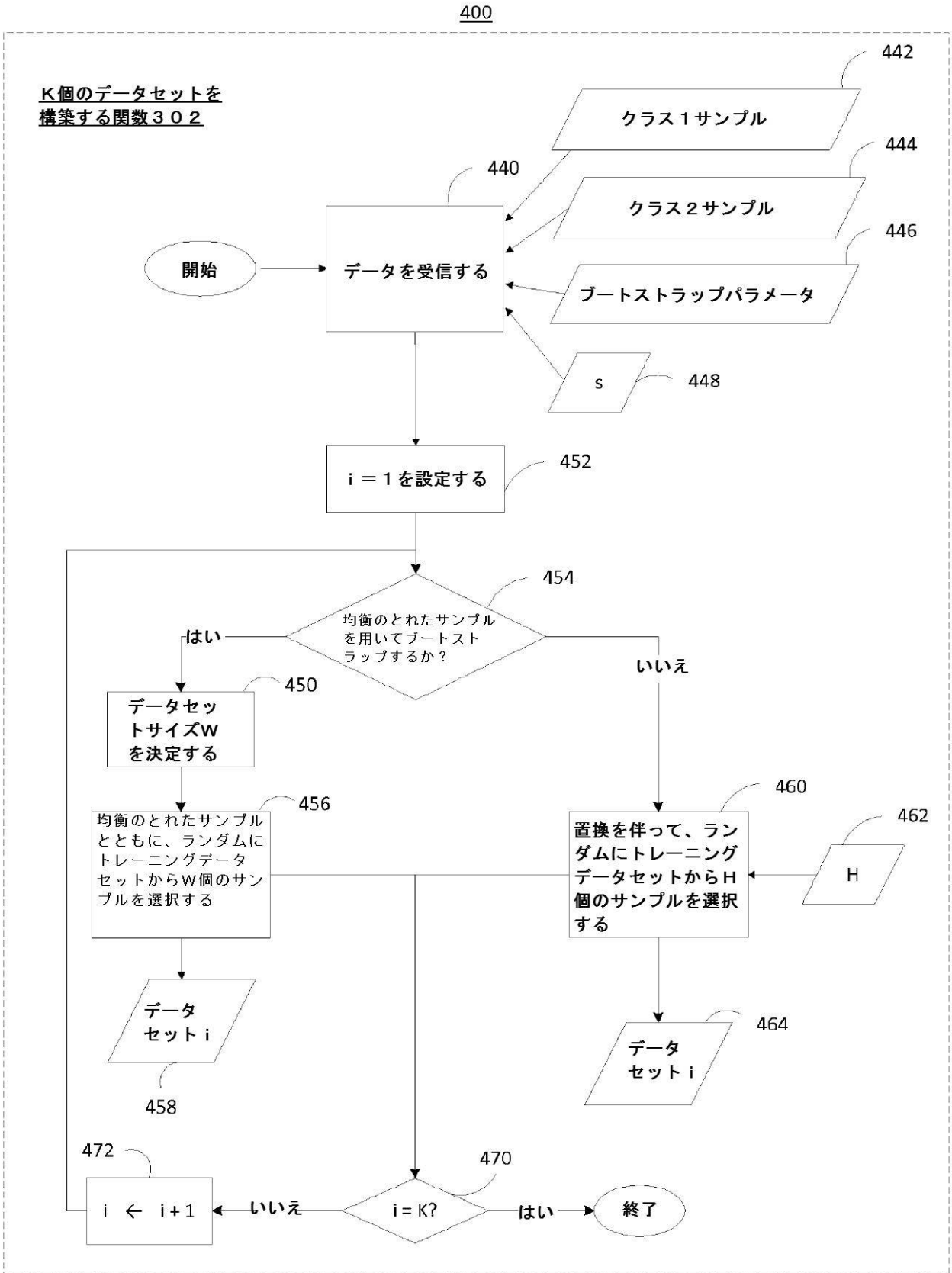


FIG. 4

【 図 5 】

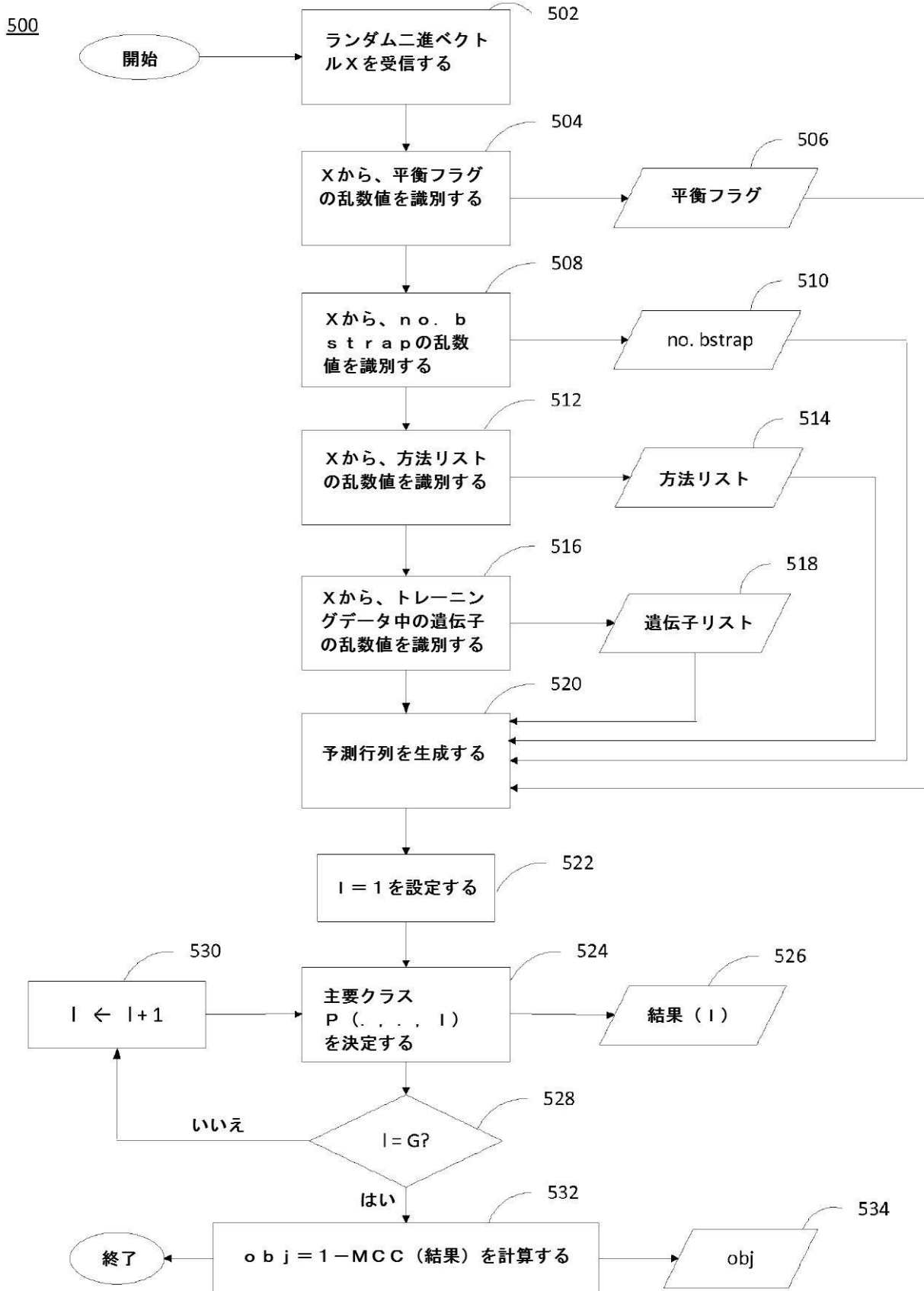


FIG. 5

【 図 6 】

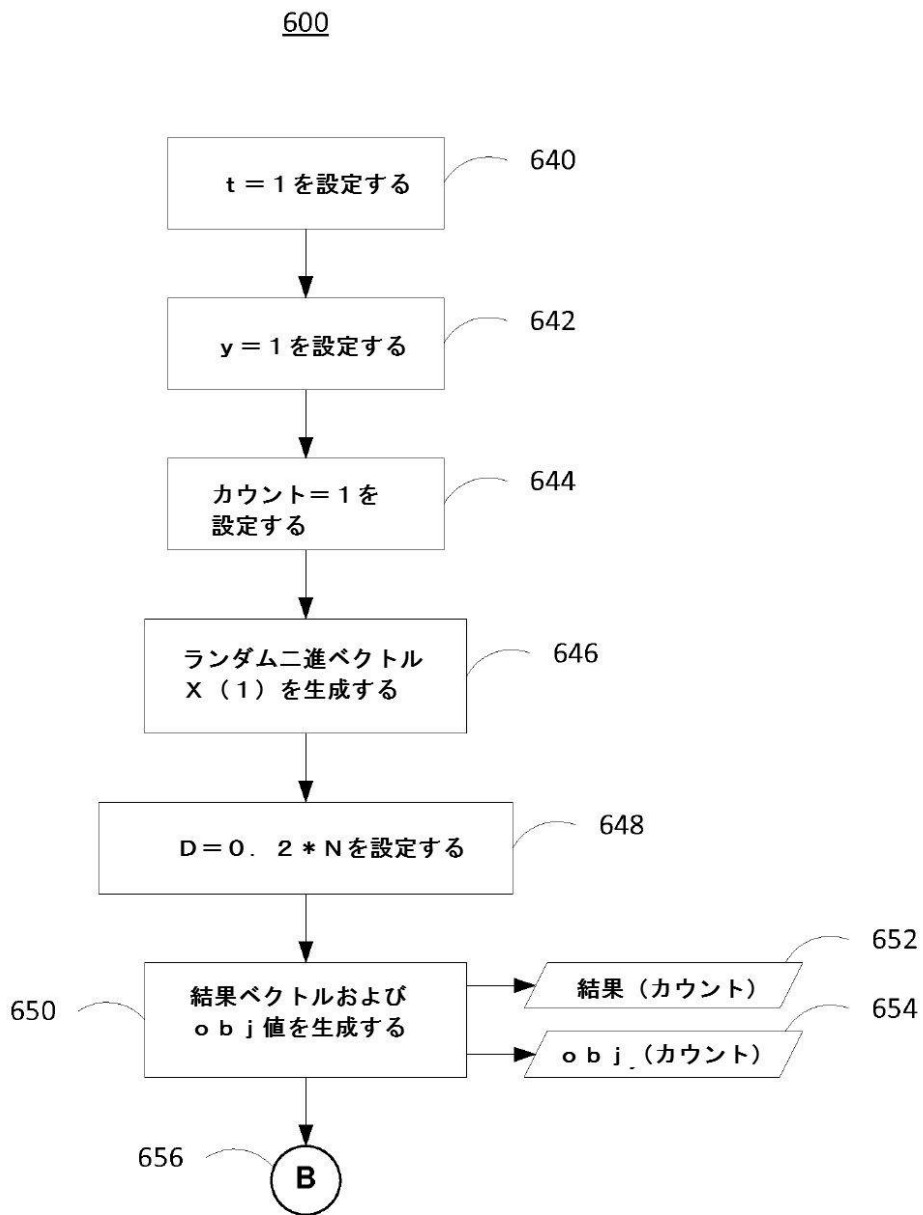


FIG. 6

【 図 7 】

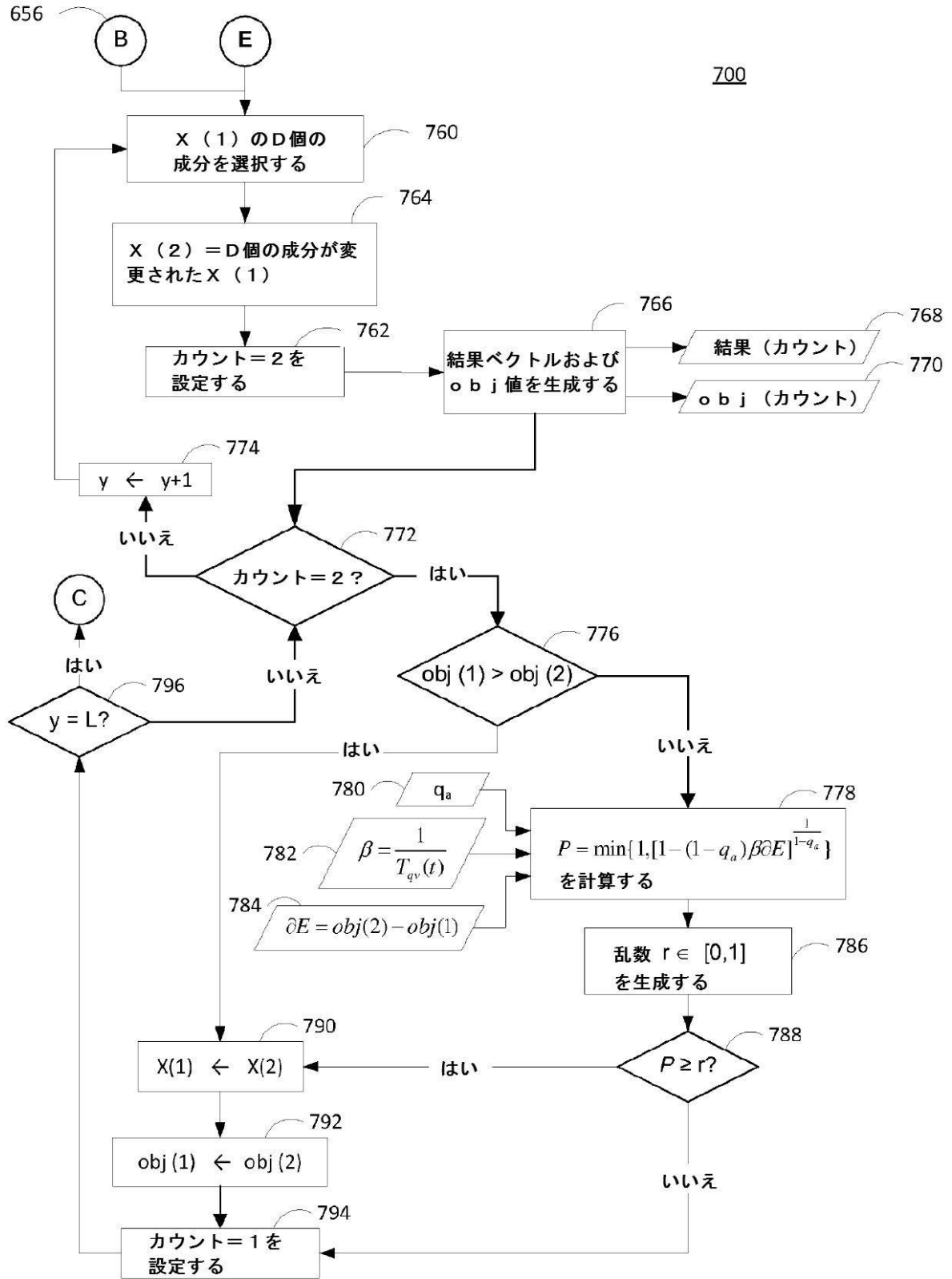
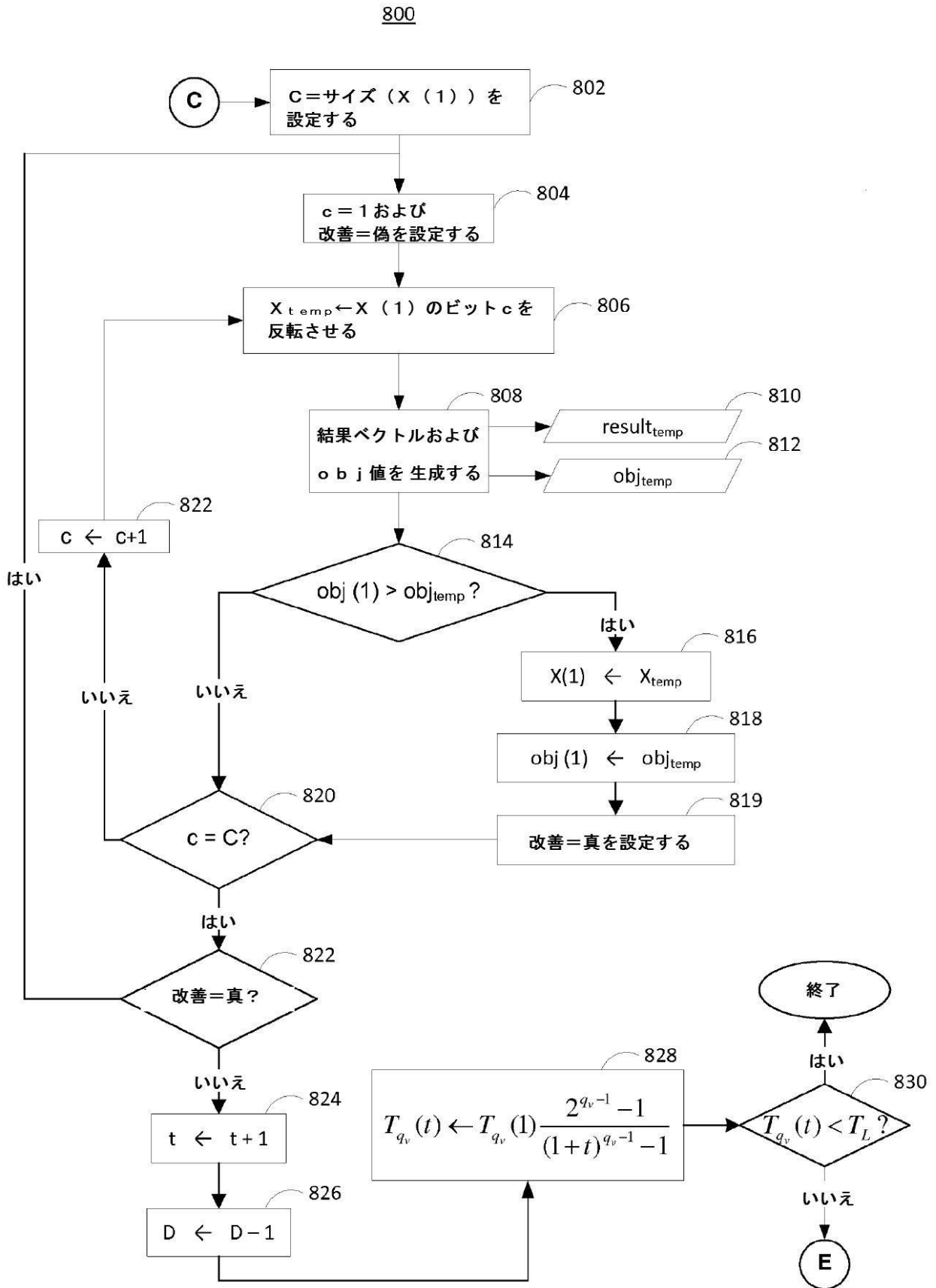


FIG. 7

【 図 8 】



【 図 9 】

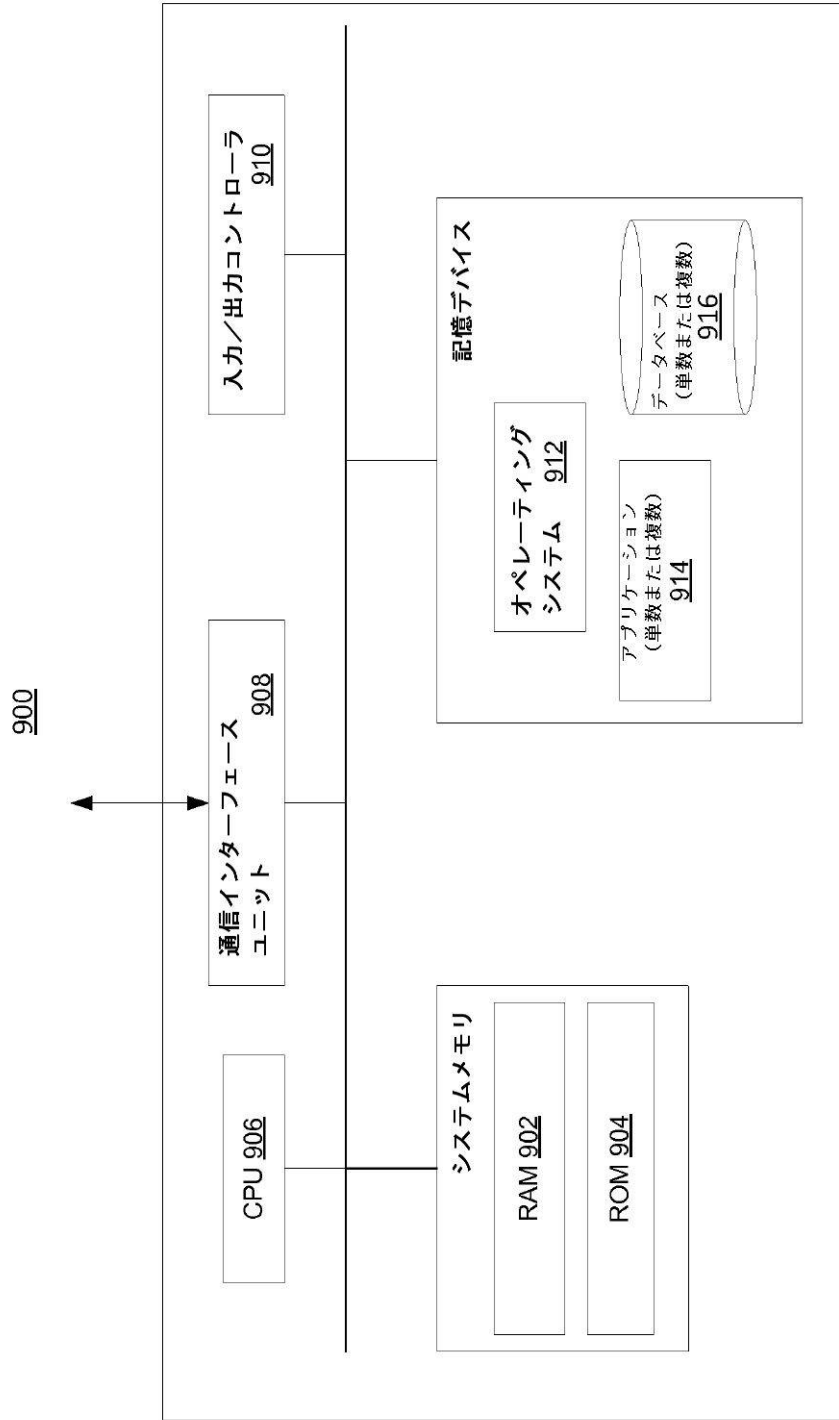


FIG. 9

【図 10】

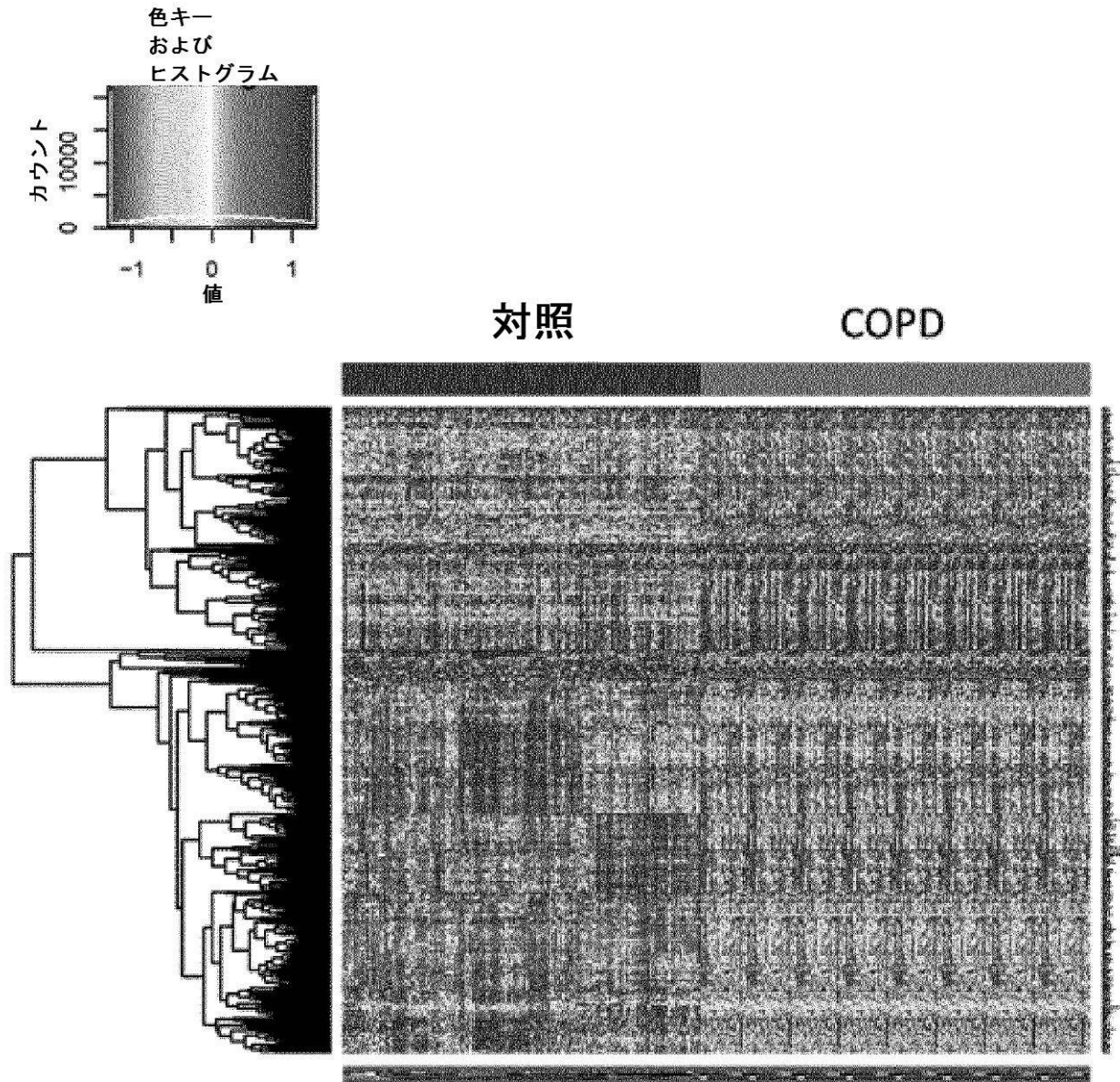


FIG. 10

【手続補正書】

【提出日】平成27年2月19日(2015.2.19)

【手続補正1】

【補正対象書類名】特許請求の範囲

【補正対象項目名】全文

【補正方法】変更

【補正の内容】

【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、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ッシュアップ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8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。

【手続補正 2】

【補正対象書類名】明細書

【補正対象項目名】0008

【補正方法】変更

【補正の内容】

【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、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、

(i i) 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、

(i i i) 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、

(i v) 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(i v)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 1 5)

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 ~ 1 3 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。

【 国際調査報告 】

INTERNATIONAL SEARCH REPORT

International application No PCT/EP2013/062982

A. CLASSIFICATION OF SUBJECT MATTER INV. G06F19/24 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G06F		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal, WPI Data		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	DANIEL GLEZ-PEÑA ET AL: "A simulated annealing-based algorithm for iterative class discovery using fuzzy logic for informative gene selection", JOURNAL OF INTEGRATED OMICS, vol. 1, no. 1, 1 February 2011 (2011-02-01), XP055084749, DOI: 10.5584/jiomics.v1i1.28 abstract page 67 page 70 page 71, column 2, paragraph 2 - paragraph 3; figure 5 ----- -/--	1-15
<input checked="" type="checkbox"/>	Further documents are listed in the continuation of Box C.	<input checked="" type="checkbox"/> See patent family annex.
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed		"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family
Date of the actual completion of the international search		Date of mailing of the international search report
5 November 2013		13/11/2013
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Schmitt, Constanze

1

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2013/062982

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	KIRKPATRICK S ET AL: "OPTIMIZATION BY SIMULATED ANNEALING", SCIENCE, AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE, WASHINGTON, DC; US, vol. 22, no. 4598, 13 May 1983 (1983-05-13), pages 671-680, XP002942889, ISSN: 0036-8075, DOI: 10.1126/SCIENCE.220.4598.671 the whole document -----	1-15
A	BLUM A ET AL: "Combining labeled and unlabeled data with co-training", 1 January 1998 (1998-01-01), PROCEEDINGS OF THE 11TH ANNUAL CONFERENCE ON COMPUTATIONAL LEARNING THEORY. COLT '98. MADISON, WI, JULY 24 - 26, 1998; [ANNUAL CONFERENCE ON COMPUTATIONAL LEARNING THEORY], NEW YORK, NY : ACM, US, PAGE(S) 92 - 100, XP007901502, ISBN: 978-0-581-13057-0 the whole document -----	1-15
A	SHIH-WEI LIN ET AL: "Parameter determination of support vector machine and feature selection using simulated annealing approach", APPLIED SOFT COMPUTING, vol. 8, no. 4, 1 September 2008 (2008-09-01), pages 1505-1512, XP055084688, ISSN: 1568-4946, DOI: 10.1016/j.asoc.2007.10.012 abstract page 1507, column 2, paragraph 4 - page 1509, column 1, paragraph 1; figure 3 -----	1-15
A	WO 2011/005893 A2 (ABBOTT LAB [US]; DEVANARAYAN VISWANATH [US]; PATTERSON MELANIE JOY [US]) 13 January 2011 (2011-01-13) page 9, paragraph 7 claim 14 -----	1-15
A	US 2012/143805 A1 (GOLD LARRY [US] ET AL) 7 June 2012 (2012-06-07) page 34, paragraph 384 page 26, paragraph 262 -----	1-15
	-/--	

1

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2013/062982

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>PENGYI YANG ET AL: "A Review of Ensemble Methods in Bioinformatics", CURRENT BIOINFORMATICS, vol. 5, no. 4, 1 December 2010 (2010-12-01), pages 296-308, XP055084854, ISSN: 1574-8936, DOI: 10.2174/157489310794072508 the whole document</p> <p style="text-align: center;">-----</p>	1-15
A	<p>Andreas Heß ET AL: "Iterative Ensemble Classification for Relational Data: A Case Study of Semantic Web Services", ECML 2004, LNAI 3201, 1 January 2004 (2004-01-01), pages 156-167, XP055084856, Retrieved from the Internet: URL:http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.658&rep=rep1&type=pdf [retrieved on 2013-10-22] page 156 - page 160</p> <p style="text-align: center;">-----</p>	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2013/062982

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2011005893 A2	13-01-2011	US 2012178637 A1 WO 2011005893 A2	12-07-2012 13-01-2011
US 2012143805 A1	07-06-2012	NONE	

フロントページの続き

(81)指定国 AP(BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), EA(AM, AZ, BY, KG, KZ, RU, TJ, TM), EP(AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OA(BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG), AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC

(71)出願人 513310690

マルティン, フロリアン

スイス国 ツェーハー - 2 0 3 4 プスー, シュマン ドゥ ロレー 1

(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

(72)発明者 シアン, ヤン

スイス国 ツェーハー - 2 0 0 0 ニューシャテル, リュ ドゥ ロシェ 2 4

(72)発明者 ヘンク, ユリア

スイス国 ツェーハー - 2 0 3 5 コルセル, グラン - リュ 3 5

(72)発明者 マルティン, フロリアン

スイス国 ツェーハー - 2 0 3 4 プスー, シュマン ドゥ ロレー 1