(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6160445号
(P6160445)
(45)発行日　平成29年7月12日(2017.7.12)　(24)登録日　平成29年6月23日(2017.6.23)

(51)Int.Cl.　F I
*G06F 17/30 (2006.01)*　G06F 17/30 210D
G06F 17/30 220Z

請求項の数 5　(全 20 頁)

(21)出願番号　特願2013-226058(P2013-226058)
(22)出願日　平成25年10月30日(2013.10.30)
(65)公開番号　特開2015-87966(P2015-87966A)
(43)公開日　平成27年5月7日(2015.5.7)
審査請求日　平成28年7月5日(2016.7.5)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　100089118
弁理士　酒井　宏明
(72)発明者　松本　和宏
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

審査官　齊藤　貴孝

最終頁に続く

(54)【発明の名称】分析装置、分析方法および分析プログラム

(57)【特許請求の範囲】
【請求項１】
入力データに対してサンプリングを実行し前記入力データから一部のデータを抽出する処理を繰り返し実行して複数のサンプリングデータを生成するサンプリング実行部と、
前記複数のサンプリングデータについてクラスタ分析を実行し、前記サンプリングデータ毎に、前記サンプリングデータに含まれるデータを異なるクラスタに分類するクラスタ分析部と、
前記複数のサンプリングデータに対する前記クラスタ分析部の複数の分類結果と前記入力データとを基にして、前記入力データに含まれるデータの所属するクラスタを予測したデータを示す予測データを複数生成するクラスタ予測部と、
前記予測データのクラスタ間距離およびクラスタ内距離を基にして、予測データ毎に評価値を算出し、パレート解となる評価値に対応する予測データを判定する判定部と、
前記パレート解となる評価値に対応する予測データを基にして、前記入力データに含まれるデータをクラスタに分類する最終クラスタ計算部と
を有することを特徴とする分析装置。
【請求項２】
前記最終クラスタ計算部は、パレート解となる評価値に対応する類似の予測データをグループ化し、同一グループに含まれる予測データを基にして、前記入力データに含まれるデータを異なるクラスタに分類する処理を、グループ毎に実行することを特徴とする請求項１に記載の分析装置。

【請求項３】
前記最終クラスタ計算部は、前記入力データに対して、ランダムにクラスタを割り当てた複数の最終クラスタデータを生成し、各最終クラスタデータと予測データとの類似度を基にして、特定の最終クラスタデータを選択することを特徴とする請求項１または２に記載の分析装置。
【請求項４】
コンピュータが実行する分析方法であって、
入力データに対してサンプリングを実行し前記入力データから一部のデータを抽出する処理を繰り返し実行して複数のサンプリングデータを生成し、
前記複数のサンプリングデータについてクラスタ分析を実行し、前記サンプリングデータ毎に、前記サンプリングデータに含まれるデータを異なるクラスタに分類し、
前記複数のサンプリングデータに対する前記クラスタ分析部の複数の分類結果と前記入力データとを基にして、前記入力データに含まれるデータの所属するクラスタを予測したデータを示す予測データを複数生成し、
前記予測データのクラスタ間距離およびクラスタ内距離を基にして、予測データ毎に評価値を算出し、パレート解となる評価値に対応する予測データを判定し、
前記パレート解となる評価値に対応する予測データを基にして、前記入力データに含まれるデータをクラスタに分類する
各処理を実行することを特徴とする分析方法。
【請求項５】
コンピュータに、
入力データに対してサンプリングを実行し前記入力データから一部のデータを抽出する処理を繰り返し実行して複数のサンプリングデータを生成し、
前記複数のサンプリングデータについてクラスタ分析を実行し、前記サンプリングデータ毎に、前記サンプリングデータに含まれるデータを異なるクラスタに分類し、
前記複数のサンプリングデータに対する前記クラスタ分析部の複数の分類結果と前記入力データとを基にして、前記入力データに含まれるデータの所属するクラスタを予測したデータを示す予測データを複数生成し、
前記予測データのクラスタ間距離およびクラスタ内距離を基にして、予測データ毎に評価値を算出し、パレート解となる評価値に対応する予測データを判定し、
前記パレート解となる評価値に対応する予測データを基にして、前記入力データに含まれるデータをクラスタに分類する
各処理を実行させることを特徴とする分析プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、分析装置等に関する。
【背景技術】
【０００２】
クラスタ分析は、データの集まりをデータ間の類似度に基づいて複数のクラスタに分類する処理である。例えば、クラスタ分析には、階層的クラスタ分析や非階層的クラスタ分析がある。
【０００３】
階層的クラスタ分析は、例えば、個々のデータを１つのクラスタとして設定し、クラスタ間の類似度を計算し、最も類似している各クラスタを併合する処理を繰り返し実行するものである。
【０００４】
非階層的クラスタ分析は、分類の状態を表す関数を使い、関数の値が最適解となるように探索を行うものである。
【先行技術文献】

【特許文献】
【０００５】
【特許文献１】特開２００７－１７９１４３号公報
【特許文献２】特開２００５－２９３０４８号公報
【発明の概要】
【発明が解決しようとする課題】
【０００６】
しかしながら、上述した従来技術では、大規模データについてクラスタ分析を実行すると時間を要するという問題がある。
【０００７】
例えば、階層的クラスタ分析および非階層的クラスタ分析はそれぞれ、小規模データ、中規模データに対してクラスタ分析を実行することを想定している。このため、現実的な計算機環境により、大規模データに対して階層的クラスタ分析や非階層的クラスタ分析を実行すると、現実的な計算時間内で計算できないことがある。
【０００８】
１つの側面では、クラスタ分析に要する時間を削減することができる分析装置、分析方法および分析プログラムを提供することを目的とする。
【課題を解決するための手段】
【０００９】
第１の案では、分析装置は、サンプリング実行部、クラスタ分析部、クラスタ予測部、判定部、最終クラスタ計算部を有する。サンプリング実行部は、入力データに対してサンプリングを実行し前記入力データから一部のデータを抽出する処理を繰り返し実行して複数のサンプリングデータを生成する。クラスタ分析部は、複数のサンプリングデータについてクラスタ分析を実行し、サンプリングデータ毎に、サンプリングデータに含まれるデータを異なるクラスタに分類する。クラスタ予測部は、複数のサンプリングデータに対するクラスタ分析部の複数の分類結果と入力データとを基にして、入力データに含まれるデータの所属するクラスタを予測したデータを示す予測データを複数生成する。判定部は、予測データのクラスタ間距離およびクラスタ内距離を基にして、予測データ毎に評価値を算出し、パレート解となる評価値に対応する予測データを判定する。最終クラスタ計算部は、パレート解となる評価値に対応する予測データを基にして、前記入力データに含まれるデータをクラスタに分類する。なお、パレート解となる予測データは、例えば、評価値が他の予測データと比較して優越するものである。
【発明の効果】
【００１０】
本発明の１実施態様によれば、クラスタ分析に要する時間を削減することができるという効果を奏する。
【図面の簡単な説明】
【００１１】
【図１】図１は、本実施例１に係る分析装置の構成を示す機能ブロック図である。
【図２】図２は、分析対象データのデータ構造の一例を示す図である。
【図３】図３は、サンプリングデータテーブルのデータ構造の一例を示す図である。
【図４】図４は、予測データテーブルのデータ構造の一例を示す図である。
【図５】図５は、評価値データテーブルのデータ構造の一例を示す図である。
【図６】図６は、中間データテーブルのデータ構造の一例を示す図である。
【図７】図７は、最終データのデータ構造の一例を示す図である。
【図８】図８は、各予測データのクラスタ内距離とクラスタ間距離との関係を示す図（１）である。
【図９】図９は、最終データ候補テーブルの一例を示す図である。
【図１０】図１０は、本実施例１にかかる分析装置の処理手順を示すフローチャートである。

【図１１】図１１は、本実施例２にかかる分析装置の構成を示す機能ブロック図である。
【図１２】図１２は、各予測データのクラスタ内距離とクラスタ間距離との関係を示す図（２）である。
【図１３】図１３は、実施例２にかかる分析装置の処理手順を示すフローチャート（１）である。
【図１４】図１４は、実施例２にかかる分析装置の処理手順を示すフローチャート（２）である。
【図１５】図１５は、分析プログラムを実行するコンピュータの一例を示す図である。
【発明を実施するための形態】
【００１２】
以下に、本願の開示する分析装置、分析方法および分析プログラムの実施例を図面に基づいて詳細に説明する。なお、この実施例によりこの発明が限定されるものではない。
【実施例１】
【００１３】
図１は、本実施例１に係る分析装置の構成を示す機能ブロック図である。図１に示すように、この分析装置１００は、通信部１１０、入力部１２０、出力部１３０、記憶部１４０、制御部１５０を有する。
【００１４】
通信部１１０は、無線または有線によってネットワークに接続し、ネットワークを介して、他の装置とデータ通信を行う処理部である。通信部１１０は、通信装置に対応する。
【００１５】
入力部１２０は、各種の情報を入力する入力装置である。入力部１２０は、例えば、キーボードやマウス、タッチパネル等に対応する。
【００１６】
出力部１３０は、制御部１５０から出力される情報を表示する表示装置である。例えば、出力部１３０は、モニタ、液晶ディスプレイ、タッチパネル等に対応する。
【００１７】
記憶部１４０は、分析対象データ１４１、サンプリングデータテーブル１４２、予測データテーブル１４３、評価値テーブル１４４、中間データテーブル１４５、最終データ１４６を有する。記憶部１４０は、例えば、ＲＡＭ（Random Access Memory）、ＲＯＭ（Read Only Memory）、フラッシュメモリ（Flash Memory）などの半導体メモリ素子や、ＨＤＤ（Hard Disk Drive）などの記憶装置に対応する。
【００１８】
分析対象データ１４１は、クラスタ分析の対象となるデータである。図２は、分析対象データのデータ構造の一例を示す図である。図２に示すように、分析対象データ１４１は、識別番号、年齢、性別、身長、体重等を有する。識別番号は、各レコードを一意に識別する情報である。年齢、性別、身長、体重は、特定の人物の年齢、性別、身長、体重をそれぞれ示す情報である。なお、図２に示す例では、性別を１または２で表す。例えば、性別「１」は、性別が男性であることを示し、性別「２」は、性別が女性であることを示す。
【００１９】
サンプリングデータテーブル１４２は、複数のサンプリングデータを有するテーブルである。各サンプリングデータは、後述するサンプリング実行部１５１によって生成される。サンプリング実行部１５１が分析対象データ１４１をサンプリングすることで、各サンプリングデータが生成される。図３は、サンプリングデータテーブルのデータ構造の一例を示す図である。図３に示すように、サンプリングデータテーブル１４２は、サンプリングデータ１４２ａ，１４２ｂ，１４２ｃを有する。図３では一例として、サンプリングデータ１４２ａ，１４２ｂ，１４２ｃを示すが、その他のサンプリングデータを含んでも良い。
【００２０】

図３において、例えば、サンプリングデータ１４２ａは、識別番号、年齢、性別、身長、体重、クラスタ番号を有する。識別番号、年齢、性別、身長、体重に関する説明は、図２で説明した、年齢、性別、身長、体重の説明と同様である。

【００２１】

予測データテーブル１４３は、複数の予測データを有するテーブルである。各予測データは、後述するクラスタ予測部１５３によって生成される。クラスタ予測部１５３が、サンプリングデータを基にして、分析対象データ１４１の各レコードのクラスタ番号を予測することで、予測データを生成する。サンプリングデータ毎に予測データが生成される。図４は、予測データテーブルのデータ構造の一例を示す図である。図４に示すように、予測データテーブル１４３は、予測データ１４３ａ，１４３ｂ，１４３ｃを有する。図４では一例として、予測データ１４３ａ，１４３ｂ，１４３ｃを示すが、その他の予測データを含んでも良い。

【００２２】

評価値データテーブル１４４は、各予測データの評価値をそれぞれ保持するテーブルである。図５は、評価値データテーブルのデータ構造の一例を示す図である。図５に示すように、この評価値データテーブル１４４は、予測データ識別情報と、評価値とを対応付ける。予測データ識別情報は、予測データを一意に識別する情報である。

【００２３】

図５において、評価値は、クラスタ間距離と、クラスタ内距離とを含む。クラスタ間距離は、異なるクラスタ間の距離を示すものである。一般的に、クラスタ間距離が大きいほど、クラスタ分析結果に対する評価が高くなる。クラスタ内距離は、クラスタの直径を示すものである。一般的に、クラスタ内距離が小さいほど、クラスタ分析結果に対する評価が高くなる。即ち、クラスタ間距離が大きいほど、また、クラスタ内距離が小さいほど、クラスタ分析結果が優れている。

【００２４】

中間データテーブル１４５は、複数の中間データを有するテーブルである。各中間データは、評価値の良い予測データに対応して作成される。図６は、中間データテーブルのデータ構造の一例を示す図である。図６に示すように、この中間データテーブル１４５は、中間データ１４５ａ，１４５ｇ，１４５ｚを有する。図６では一例として、中間データ１４５ａ，１４５ｇ，１４５ｚを示すが、その他の中間データを含んでも良い。

【００２５】

図６において、中間データは、識別番号と、各クラスタ番号とを対応付ける。識別番号は、分析対象データ１４１の識別番号に対応する。例えば、中間データ１４５ａの１段目では、識別番号１００１に対応するレコードが、クラスタ番号「１」に分類されることを示す。ここで、識別番号「１００１」に対応するレコードは、図２に示した分析対象データ１４１の識別番号「１００１」に対応するレコードに対応する。従って、図２に示した分析対象データ１４１の識別番号「１００１」に対応するレコードが、クラスタ番号「１」のクラスタに属していることを示す。

【００２６】

最終データ１４６は、分析対象データ１４１の最終的なクラスタ分析結果を示す。図７は、最終データのデータ構造の一例を示す図である。図７に示すように、この最終データ１４６は、識別番号と、各クラスタ番号とを対応付ける。識別番号は、分析対象データ１４１の識別番号に対応する。例えば、最終データ１４６の１段目では、識別番号１００１に対応するレコードが、クラスタ番号「１」に分類されることを示す。

【００２７】

図１の説明に戻る。制御部１５０は、サンプリング実行部１５１、クラスタ分析部１５２、クラスタ予測部１５３、判定部１５４、最終クラスタ計算部１５５を有する。制御部１５０は、例えば、ＡＳＩＣ（Application Specific Integrated Circuit）や、ＦＰＧＡ（Field Programmable Gate Array）などの集積装置に対応する。また、制御部１５０は、例えば、ＣＰＵ（Central Processing Unit）やＭＰＵ（Micro Processing

Unit）等の電子回路に対応する。

【００２８】

サンプリング実行部１５１は、分析対象データ１４１に対してサンプリングを複数回繰り返し実行することで、複数のサンプリングデータを生成する処理部である。サンプリング実行部１５１は、生成した各サンプリングデータを、サンプリングデータテーブル１４２に格納する。

【００２９】

例えば、サンプリング実行部１５１は、入力部１２０を介して、計算回数およびサンプリング件数を取得し、取得した計算回数だけ、サンプリングを行う。また、サンプリング実行部１５１は、サンプリングを実行する度に、サンプリング間隔を変更しても良い。また、ランダムサンプリングを行っても良い。

【００３０】

サンプリング実行部１５１は、サンプリングデータのレコードの件数を、入力部１２０から取得したサンプリング件数に合わせる。例えば、指定されたサンプリング件数がＮ２件の場合には、各サンプリングデータの件数をそれぞれＮ２件とする。例えば、分析対象データのレコードの件数をＮ１件とすると、Ｎ１とＮ２との大小関係は「Ｎ１＞Ｎ２」となる。

【００３１】

クラスタ分析部１５２は、サンプリングデータテーブル１４２に格納された各サンプリングデータを取得し、各サンプリングデータをクラスタ分析する処理部である。クラスタ分析部１５２は、クラスタ分析結果に応じて、サンプリングデータの各レコードについてクラスタ番号を割り当てる。

【００３２】

図３に示したサンプリングデータテーブル１４２を例にして説明を行う。クラスタ分析部１５２は、まず、サンプリングデータ１４２ａに対してクラスタ分析を行い、サンプリングデータ１４２ａの各レコードを複数のクラスタに分類し、分類結果に応じて、クラスタ番号を割り振る。クラスタ分析部１５２は、サンプリングデータ１４２ｂ，１４２ｃについても同様に、クラスタ分析を行って、各レコードを、複数のクラスタに分類し、分類結果に応じて、クラスタ番号を割り振る。クラスタ分析部１５２が分類するクラスタの数は、予め設定されているものとする。

【００３３】

クラスタ分析部１５２が行うクラスタ分析は、階層的クラスタ分析でも良いし、非階層的クラスタ分析でもよい。ここでは一例として、クラスタ分析部１５２が、階層的クラスタ分析を実行する場合について説明する。

【００３４】

クラスタ分析部１５２が、階層的クラスタ分析を行う場合には、まず、個々のデータを１つのクラスタとして設定し、クラスタ間の類似度を計算する。クラスタ分析部１５２は、最も類似しているクラスタを併合する。クラスタ分析部１５２は、予め設定されたクラスタの数と同数になるまで、上記処理を繰り返し実行する。

【００３５】

例えば、クラスタ分析部１５２は、各クラスタの組み合わせについて、クラスタ間のユークリッド距離を算出し、ユークリッド距離が最小となる各クラスタの組みを、合併する。この場合には、クラスタ間のユークリッド距離が上記クラスタ間の類似度に対応し、ユークリッド距離が短いほど、類似度が高い。

【００３６】

クラスタ予測部１５３は、サンプリングデータテーブル１４２のサンプリングデータのクラスタ分析結果に基づいて、分析対象データ１４１の各レコードのクラスタ番号を予測し、予測データテーブル１４３を生成する処理部である。クラスタ予測部１５３は、サンプリングデータテーブル１４２に含まれるサンプリングデータの数だけ、予測データを生成し、生成した予測データを予測データテーブル１４３に登録する。

【００３７】
例えば、クラスタ予測部１５３は、サンプリングデータテーブル１４２のサンプリングデータ１４２ａを基にして、予測データ１４３ａを生成する。クラスタ予測部１５３は、サンプリングデータ１４２ｂを基にして、予測データ１４３ｂを生成する。クラスタ予測部１５３は、サンプリングデータ１４２ｃを基にして、予測データ１４３ｃを生成する。クラスタ予測部１５３は、サンプリングデータがＮ個存在する場合には、予測データをＮ個作成する。
【００３８】
ここで、クラスタ予測部１５３が、サンプリングデータ１４２ａを基にして、予測データ１４３ａを生成する場合の処理の一例について説明する。まず、クラスタ予測部１５３は、サンプリングデータ１４２含まれる識別番号と、クラスタ番号との関係を、そのまま、予測データ１４３ａに設定する。
【００３９】
例えば、クラスタ予測部１５３は、サンプリングデータ１４２ａに識別番号「１００１」のレコードのクラスタ番号が「１」の場合には、予測データ１４３ａの識別番号「１００１」のクラスタ番号を「１」に設定する。同様に、クラスタ予測部１５３は、サンプリングデータ１４２ａに存在する全ての識別番号とクラスタ番号との関係を、予測データ１４３ａに設定する。
【００４０】
続いて、クラスタ予測部１５３は、上記処理を行った結果、クラスタ番号が未設定となるレコードについて下記の処理を行う。まず、クラスタ予測部１５３は、各クラスタに分類されたレコードから、代表レコードを検出する。例えば、クラスタ番号「１」のレコードのうち、平均的な数値を有するレコードを代表レコードとして検出する。クラスタ予測部１５３は、他のクラスタ番号に対応する代表レコードも同様にして検出する。
【００４１】
クラスタ予測部１５３は、クラスタ番号が未設定のレコードと、各代表レコードとのユークリッド距離を計算し、ユークリッド距離が最小となる組み合わせを特定する。クラスタ予測部１５３は、特定した組の代表レコードのクラスタ番号を、該当するレコードのクラスタ番号に設定する。
【００４２】
例えば、クラスタ番号が未設定のレコードと、各代表レコードとのユークリッド距離を算出し、未設定のレコードと、クラスタ番号「１」の代表レコードとのユーグリッド距離が最小の場合には、該当するレコードのクラスタ番号を「１」に設定する。クラスタ予測部１５３は、未設定のレコードについて、上記処理を繰り返し実行することで、予測データテーブル１４３を生成する。
【００４３】
判定部１５４は、予測データテーブル１４３を基にして、評価値データテーブル１４４を生成する処理部である。評価部１５４は、予測データテーブル１４３に含まれる予測データ毎に評価値を算出する。
【００４４】
判定部１５４は、予測データ毎にクラスタ間距離およびクラスタ内距離を算出し、クラスタ間距離およびクラスタ内距離を予測データの評価値とする。予測データのクラスタ間距離を算出する処理の一例について説明する。ここでは、クラスタ番号「１～３」のクラスタが存在するものとする。判定部１５４は、クラスタ番号「１」に属する第１代表レコードと、クラスタ番号「２」に属する第２代表レコードと、クラスタ番号「３」に属する第３代表レコードとを検出する。代表レコードを検出する処理の一例は、例えば、同一のクラスタ番号に属するレコードのうち、平均的な数値を有するレコードを代表レコードとして検出する。
【００４５】
判定部１５４は、第１代表レコードと、第２代表レコードとのユークリッド距離を算出

し、第１代表レコードと第３代表レコードとのユークリッド距離を算出する。判定部１５４は、算出した各ユークリッド距離を平均したユークリッド距離を、予測データのクラスタ間距離とする。

【００４６】

例えば、第１代表レコードの年齢、性別、身長、体重の値をそれぞれ、ａ１、ａ２、ａ３、ａ４とする。第２代表レコードの年齢、性別、身長、体重の値をそれぞれ、ｂ１、ｂ２、ｂ３、ｂ４とする。第３代表レコードの年齢、性別、身長、体重の値をそれぞれ、ｃ１、ｃ２、ｃ３、ｃ４とする。この場合には、第１代表レコードと、第２代表レコードとのユークリッド距離Ｘ１は、式（１）で計算され、第１代表レコードと、第３代表レコードとのユークリッド距離Ｘ２は、式（２）で計算される。この場合には、予測データのクラスタ間距離は式（３）に示すものとなる。

【００４７】

ユーグリット距離Ｘ１＝（（ａ１－ｂ１）$^{2}$＋（ａ２－ｂ２）$^{2}$＋（ａ３－ｂ３）$^{2}$＋（ａ４－ｂ４）$^{2}$）$^{1/2}$・・・（１）

【００４８】

ユーグリット距離Ｘ２＝（（ａ１－ｃ１）$^{2}$＋（ａ２－ｃ２）$^{2}$＋（ａ３－ｃ３）$^{2}$＋（ａ４－ｃ４）$^{2}$）$^{1/2}$・・・（２）

【００４９】

クラスタ間距離＝（Ｘ１＋Ｘ２）／２・・・（３）

【００５０】

続いて、クラスタ内距離を算出する処理について説明する。まず、判定部１５４は、同一のクラスタ番号に属する各レコード間のユークリッド距離をそれぞれ算出する。そして、判定部１５４は、算出したユークリッド距離を平均したユークリッド距離を、予測データのクラスタ内距離とする。判定部１５４は、各クラスタ番号のクラスタに対応するクラスタ内距離を平均することで、予測データのクラスタ内距離を算出する。

【００５１】

例えば、クラスタ番号「１～３」のクラスタが存在する場合には、判定部１５４は、各クラスタ番号「１～３」のクラスタ内距離をそれぞれ算出する。判定部１５４は、各クラスタ番号「１～３」のクラスタ内距離を平均することで、予測データのクラスタ内距離を算出する。

【００５２】

例えば、クラスタ番号「１」のクラスタ内距離を算出する例について説明する。クラスタ内に３つの第１レコード、第２レコード、第３レコードが存在するものとする。例えば、第１レコードの年齢、性別、身長、体重の値をそれぞれ、ｄ１、ｄ２、ｄ３、ｄ４とする。第２レコードの年齢、性別、身長、体重の値をそれぞれ、ｅ１、ｅ２、ｅ３、ｅ４とする。第３代表レコードの年齢、性別、身長、体重の値をそれぞれ、ｆ１、ｆ２、ｆ３、ｆ４とする。この場合には、第１レコードと、第２レコードとのユークリッド距離Ｙ１は、式（４）で計算され、第１レコードと、第３レコードとのユークリッド距離Ｙ２は、式（５）で計算される。この場合には、クラスタ番号「１」のクラスタのクラスタ内距離は式（６）に示すものとなる。

【００５３】

ユーグリット距離Ｙ１＝（（ｄ１－ｅ１）$^{2}$＋（ｄ２－ｅ２）$^{2}$＋（ｄ３－ｅ３）$^{2}$＋（ｄ４－ｅ４）$^{2}$）$^{1/2}$・・・（４）

【００５４】

ユーグリット距離Ｙ２＝（（ｄ１－ｆ１）$^{2}$＋（ｄ２－ｆ２）$^{2}$＋（ｄ３－ｆ３）$^{2}$＋（ｄ４－ｆ４）$^{2}$）$^{1/2}$・・・（５）

【００５５】

クラスタ内距離＝（Ｙ１＋Ｙ２）／２・・・（６）

【００５６】

判定部１５４は、他のクラスタについても同様にクラスタ内距離を算出し、各クラスタ

のクラスタ内距離を平均することで、予測データのクラスタ内距離を算出する。

【００５７】

判定部１５４は、予測データテーブル１４３に含まれる予測データ毎に上記処理を実行することで、各予測データの評価値を算出し、評価値データテーブル１４４を生成する。

【００５８】

最終クラスタ計算部１５５は、分析対象データ１４１の最終的なクラスタ分析結果となる最終データ１４６を生成する処理部である。最終クラスタ計算部１５５は、評価値データテーブル１４４から中間データテーブル１４５を生成する処理を行った後に、中間データテーブル１４５を基にして、最終データ１４６を生成する。

【００５９】

最終クラスタ計算部１５５が、評価値データテーブル１４４から中間データテーブル１４５を生成する処理の一例について説明する。最終クラスタ計算部１５５は、評価値データテーブル１４４の予測データ毎の評価値を比較して、パレート解となる予測データを特定し、特定したパレート解となる予測データを、中間データテーブル１４５に設定する。例えば、パレート解となる予測データは、一つ以上の項目について他の予測データよりも優れているものとなる。

【００６０】

図８は、各予測データのクラスタ内距離とクラスタ間距離との関係を示す図（１）である。図８において、縦軸はクラスタ内距離を示し、横軸はクラスタ間距離を示す。一般的に、クラスタ間距離が大きいほど、また、クラスタ内距離が小さいほど、予測データは、良い予測データであると言える。このため、図８に示す例では、最終クラスタ計算部１５５は、予測データ１４３ａ，１４３ｇ，１４３ｚを、パレート解として特定する。

【００６１】

続いて、最終クラスタ計算部１５５が、中間データテーブル１４５から最終データ１４６を生成する処理について説明する。まず、最終クラスタ計算部１５５は、最終データ候補テーブルを生成する。図９は、最終データ候補テーブルの一例を示す図である。図９に示すように、この最終データ候補テーブル１０は、最終データ候補１０ａ，１０ｂ，１０ｃを有する。ここでは一例として、最終データ候補１０ａ，１０ｂ，１０ｃを示すが、これ以外に、最終データ候補を含んでいても良い。

【００６２】

最終クラスタ計算部１５５は、最終データ候補１０ａ，１０ｂ，１０ｃの各クラスタ番号を０の初期値に設定する。そして、最終クラスタ計算部１５５は、各識別番号の各クラスタ番号の値のいずれか一つが「１」となるように、ランダムに「１」を割り振る。例えば、図９に示す例では、最終データ候補１０ａの識別番号「１００１」に対してランダムに「１」を割り振ることで、クラスタ番号「１」に対応するものが「１」に設定され、その他のクラスタ番号については「０」が設定される。

【００６３】

最終クラスタ計算部１５５は、最終データ候補テーブル１０の各最終データ候補１０ａ，１０ｂ，１０ｃと、中間データテーブル１４５の各中間データとの類似度を計算し、最も類似度の高い最終データ候補を、最終データ１４６として特定する。

【００６４】

最終クラスタ計算部１５５は、中間データの識別番号および識別番号に対応するクラスタ番号と、最終データ候補の識別番号および識別番号に対応するクラスタ番号とを比較し、一致する数を計数する。最終クラスタ計算部１５５は、一致する数を、全レコード数で除算することで、類似度を算出する。以下の説明では、一致する数を、一致数と表記する。

【００６５】

例えば、最終クラスタ計算部１５５が、最終データ候補１０ａの類似度を算出する場合について説明する。最終クラスタ計算部１５５は、最終データ候補１０ａと中間データ１４５ａとを比較し、一致数が「Ｌ１」であり、最終データ候補１０ａの全レコード数が「

Ｍ１」の場合には、最終データ候補１０ａと中間データ１４５ａとの類似度は「Ｌ１／Ｍ１」となる。最終クラスタ計算部１５５は、最終データ候補１０ａと中間データ１４５ｇとを比較し、一致数が「Ｌ２」であり、最終データ候補１０ａの全レコード数が「Ｍ２」の場合には、最終データ候補１０ａと中間データ１４５ａとの類似度は「Ｌ２／Ｍ２」となる。最終クラスタ計算部１５５は、最終データ候補１０ａと中間データ１４５ｚとを比較し、一致数が「Ｌ３」であり、最終データ候補１０ａの全レコード数が「Ｍ３」の場合には、最終データ候補１０ａと中間データ１４５ｚとの類似度は「Ｌ３／Ｍ３」となる。この場合には、最終クラスタ計算部１５５は、最終データ候補１０ａの類似度を「Ｌ１／Ｍ１＋Ｌ２／Ｍ２＋Ｌ３／Ｍ３」と特定する。

【００６６】

最終クラスタ計算部１５５は、最終データ候補１０ｂ，１０ｃに関しても、最終データ候補１０ａと同様にして、類似度を算出する。最終クラスタ計算部１５５は、最終データ候補１０ａの類似度、最終データ候補１０ｂの類似度、最終データ候補１０ｃの類似度を比較し、類似度が最大となる最終データ候補を特定する。最終クラスタ計算部１５５は、特定した最終データ候補を、最終データ１４６として設定する。最終クラスタ計算部１５５は、最終データ１４６を、出力部１３０に出力しても良い。

【００６７】

次に、本実施例１にかかる分析装置１００の処理手順について説明する。図１０は、本実施例１にかかる分析装置の処理手順を示すフローチャートである。図１０に示すように、分析装置１００は、分析対象データ１４１を受け付ける（ステップＳ１０１）。また、分析装置１００は、繰り返し計算回数を受け付け（ステップＳ１０２）、サンプリング件数を受け付ける。また、カウント値を初期化する（ステップＳ１０３）。分析装置１００は、カウント値に１を加算する（ステップＳ１０４）。カウント値の初期値を０とする。

【００６８】

分析装置１００は、分析対象データ１４１をサンプリングし、サンプリングデータを生成する（ステップＳ１０５）。各サンプリングデータは、サンプリングデータテーブル１４２に格納される。分析装置１００は、サンプリングデータに対してクラスタ分析処理を実行し、各々のレコードに対してクラスタ番号を割り振る（ステップＳ１０６）。

【００６９】

分析装置１００は、クラスタ番号を割り振ったサンプリングデータと分析対象データ１４１とを比較して、分析対象データ１４１に含まれる各々のレコードに対してクラスタ番号を割り振ることで予測データを生成する（ステップＳ１０７）。各予測データは、予測データテーブル１４３に格納される。

【００７０】

分析装置１００は、予測データを基にして、クラスタ内距離およびクラスタ間距離を算出し、評価値データテーブル１４４を生成する（ステップＳ１０８）。分析装置１００は、繰り返しの計算回数がカウント値未満であるか否かを判定する（ステップＳ１０９）。分析装置１００は、繰り返しの計算回数がカウント値未満の場合には（ステップＳ１０９０，Ｙｅｓ）、ステップＳ１０４に移行する。

【００７１】

一方、分析装置１００は、繰り返しの計算回数がカウント値以上である場合には（ステップＳ１０９，Ｎｏ）、パレート解に対応する予測データを選択して、中間データテーブル１４５を作成する（ステップＳ１１０）。

【００７２】

分析装置１００は、ランダムにクラスタ番号を割り振った複数の最終データ候補を生成する（ステップＳ１１１）。ステップＳ１１１において、分析装置１００は、類似度が大きくなるようにクラスタ番号を割り振る。例えば、分析装置１００は、ランダムにクラスタ番号を割り振り、類似度を計算する。そして、分析装置１００は、類似度が大きい、クラスタ番号の割り振りを少し変更して、類似度が大きくなるか、試行する処理を利用者が設定した回数繰り返す。

【００７３】
分析装置１００は、中間データと各最終データ候補とを比較して、類似度が最大となる最終データ候補を判定する（ステップＳ１１２）。ステップＳ１１２で判定した類似度が最大となる最終データ候補が、最終データ１４６となる。分析装置１００は、判定結果を出力する（ステップＳ１１３）。
【００７４】
次に、本実施例１にかかる分析装置１００の効果について説明する。分析装置１００は分析対象データ１４１から抽出したサンプリングデータをクラスタ分析し、サンプリングデータのクラスタ分析結果を基にして、分析対象データの各データが属するクラスタを予測した複数の予測データを生成する。そして、分析装置１００は、複数の予測データのうち、評価値のよい予測データのクラスタ分類結果を用いて、分析対象データ１４１の最終的なクラスタ分類結果を特定する。これにより、分析装置１００によれば、クラスタ分析に要する時間を削減することができる。
【００７５】
また、現実的な計算機で、現実的な時間内に計算できない、大規模なデータに対して、現実的な計算機で、現実的な時間内に、クラスタ分析を実行することができる。
【実施例２】
【００７６】
図１１は、本実施例２にかかる分析装置の構成を示す機能ブロック図である。図１１に示すように、この分析装置２００は、通信部２１０、入力部２２０、出力部２３０、記憶部２４０、制御部２５０を有する。
【００７７】
通信部２１０、入力部２２０、出力部２３０に関する説明は、図１に示した、通信部１１０、入力部１２０、出力部１３０に関する説明と同様である。
【００７８】
記憶部２４０は、分析対象データ２４１、サンプリングデータテーブル２４２、予測データテーブル２４３、評価値データテーブル２４４、中間データテーブル２４５、最終データ２４６を有する。記憶部２４０は、例えば、ＲＡＭ、ＲＯＭ、フラッシュメモリなどの半導体メモリ素子や、ＨＤＤなどの記憶装置に対応する。
【００７９】
分析対象データ２４１は、クラスタ分析の対象となるデータである。分析対象データ２４１のデータ構造は、図２に示した分析対象データ１４１のデータ構造と同様である。
【００８０】
サンプリングデータテーブル２４２は、複数のサンプリングデータを有するテーブルである。各サンプリングデータは、後述するサンプリング実行部２５１によって生成される。サンプリング実行部２５１が分析対象データ２４１をサンプリングすることで、各サンプリングデータが生成される。サンプリングデータテーブル２４２のデータ構造は、図３に示したサンプリングデータテーブル１４２のデータ構造と同様である。
【００８１】
予測データテーブル２４３は、複数の予測データを有するテーブルである。各予測データは、後述するクラスタ予測部２５３によって生成される。クラスタ予測部２５３が、サンプリングデータを基にして、分析対象データ２４１の各レコードのクラスタ番号を予測することで、予測データを生成する。サンプリングデータ毎に予測データが生成される。予測データテーブル２４３のデータ構造は、図４に示した予測データテーブル１４３のデータ構造と同様である。
【００８２】
評価値データテーブル２４４は、各予測データの評価値をそれぞれ保持するテーブルである。評価値データテーブル２４４のデータ構造は、図５に示した評価値データテーブル１４４のデータ構造と同様である。
【００８３】

中間データテーブル２４５は、複数の中間データを有するテーブルである。各中間データは、評価値の良い予測データに対応して作成される。中間データテーブル２４５のデータ構造は、図６に示した中間データテーブル１４５のデータ構造と同様である。

【００８４】

最終データ２４６ａ，２４６ｂ，２４６ｃは、分析対象データ２４１の最終的なクラスタ分析結果を示す。各最終データ２４６ａ，２４６ｂ，２４６ｃのデータ構造は、図７に示した最終データ１４６のデータ構造と同様である。

【００８５】

図１１の説明に戻る。制御部２５０は、サンプリング実行部２５１、クラスタ分析部２５２、クラスタ予測部２５３、判定部２５４、最終クラスタ計算部２５５を有する。制御部２５０は、例えば、ＡＳＩＣや、ＦＰＧＡなどの集積装置に対応する。また、制御部２５０は、例えば、ＣＰＵやＭＰＵ等の電子回路に対応する。

【００８６】

サンプリング実行部２５１は、分析対象データ２４１に対してサンプリングを複数回繰り返し実行することで、複数のサンプリングデータを生成する処理部である。サンプリング実行部２５１は、生成した各サンプリングデータを、サンプリングデータテーブル２４２に格納する。サンプリング実行部２５１の具体的な処理は、図１に示したサンプリング実行部１５１と同様である。

【００８７】

クラスタ分析部２５２は、サンプリングデータテーブル２４２に格納された各サンプリングデータを取得し、各サンプリングデータをクラスタ分析する処理部である。クラスタ分析部２５２は、クラスタ分析結果に応じて、サンプリングデータの各レコードについてクラスタ番号を割り当てる。クラスタ分析部２５２の具体的な処理は、図１に示したクラスタ分析部１５２と同様である。

【００８８】

クラスタ予測部２５３は、サンプリングデータテーブル２４２のサンプリングデータのクラスタ分析結果に基づいて、分析対象データ２４１の各レコードのクラスタ番号を予測し、予測データテーブル２４３を生成する処理部である。クラスタ予測部２５３の具体的な処理は、図１に示したクラスタ予測部１５３と同様である。

【００８９】

判定部２５４は、予測データテーブル２４３を基にして、評価値データテーブル２４４を生成する処理部である。判定部２５４は、予測データテーブル２４３に含まれる予測データ毎に評価値を算出する。例えば、判定部２５４は、予測データ毎にクラスタ間距離およびクラスタ内距離を算出し、クラスタ間距離およびクラスタ内距離を予測データの評価値とする。

【００９０】

最終クラスタ計算部２５５は、分析対象データ２４１の最終的なクラスタ分析結果となる最終データ２４６を生成する処理部である。最終クラスタ計算部２５５は、評価値データテーブル２４４から中間データテーブル２４５を生成する処理を行った後に、中間データテーブル２４５を基にして、最終データ２４６ａ，２４６ｂ，２４６ｃを生成する。

【００９１】

最終クラスタ計算部２５５が、評価値データテーブル２４４から中間データテーブル２４５を生成する処理について説明する。最終クラスタ計算部２５５は、評価データテーブル２４４の予測データ毎の評価値を比較し、パレート解となる予測データを特定し、特定したパレート解となる予測データを、中間データテーブル２４５に設定する。

【００９２】

図１２は、各予測データのクラスタ内距離とクラスタ間距離との関係を示す図（２）である。図１２において、縦軸はクラスタ内距離を示し、横軸はクラスタ間距離を示す。一般的に、クラスタ間距離が大きいほど、また、クラスタ内距離が小さいほど、中間データは、良い予測データであると言える。このため、図１２に示す例では、最終クラスタ計算

部２５５は、中間データ２４３ａ，２４３ｃ，２４３ｆ，２４３ｇ，２４３ｚを、パレート解として特定する。

【００９３】

続いて、最終クラスタ計算部２５５が、中間データテーブル２４５から最終データ２４６を生成する処理について説明する。最終クラスタ計算部２５５は、中間データテーブル２４５の各予測データの評価値を比較して、類似する予測データ同士を同一グループに分類する処理を行う。例えば、最終クラスタ計算部２５５は、各予測データのクラスタ間距離の差分が閾値未満となり、かつ、各予測データのクラスタ内距離の差分が閾値未満となる予測データを、同一のグループに分類する。

【００９４】

図１２に示す例では、最終クラスタ計算部２５５は、予測データ２４３ａ，２４３ｃをグループ５０ａに分類し、予測データ２４３ｆ，２４３ｇをグループ５０ｂに分類し、予測データ２４３ｉ，２４３ｚをグループ５０ｃに分類する。最終クラスタ計算部２５５は、分類したグループ毎に、最終データ２４６を生成する。

【００９５】

例えば、最終クラスタ計算部２５５は、グループ５０ａに含まれる予測データ２４３ａ，２４３ｃを基にして、最終データ２４６ａを生成する。最終クラスタ計算部２５５は、グループ５０ｂに含まれる予測データ２４３ｆ，２４３ｇを基にして、最終データ２４６ｂを生成する。最終クラスタ計算部２５５は、グループ５０ｃに含まれる予測データ２４３ｉ，２４３ｚを基にして、最終データ２４６ｃを生成する。

【００９６】

最終クラスタ計算部２５５が、予測データを基にして、最終データを特定する処理は、図１の最終クラスタ計算部１５５が、中間データテーブル１４５の予測データを基にして、最終データを特定する処理と同様である。図１２に示す例では、グループ５０ａ，５０ｂ，５０ｃについて、最終データが特定され、最終データ２４６ａ，２４６ｂ，２４６ｃが生成される。

【００９７】

次に、本実施例２に係る分析装置２００の処理手順について説明する。図１３および図１４は、実施例２にかかる分析装置の処理手順を示すフローチャートである。図１３に示すように、分析装置２００は、分析対象データ２４１を受け付ける（ステップＳ２０１）。また、分析装置２００は、繰り返し計算回数を受け付け（ステップＳ２０２）、サンプリング件数を受け付ける。また、カウント値を初期化する（ステップＳ２０３）。分析装置２００は、カウント値に１を加算する（ステップＳ２０４）。カウント値の初期値を０とする。

【００９８】

分析装置２００は、分析対象データ２４１をサンプリングし、サンプリングデータを生成する（ステップＳ２０５）。各サンプリングデータは、サンプリングデータテーブル２４２に格納される。分析装置２００は、サンプリングデータに対してクラスタ分析処理を実行し、各々のレコードに対してクラスタ番号を割り振る（ステップＳ２０６）。

【００９９】

分析装置２００は、クラスタ番号を割り振ったサンプリングデータと分析対象データ２４１とを比較して、分析対象データ２４１に含まれる各々のレコードに対してクラスタ番号を割り振ることで予測データを生成する（ステップＳ２０７）。各予測データは、予測データテーブル２４３に格納される。

【０１００】

分析装置２００は、予測データを基にして、クラスタ内距離およびクラスタ間距離を算出し、評価値データテーブル２４４を生成する（ステップＳ２０８）。分析装置２００は、繰り返しの計算回数がカウント値未満であるか否かを判定する（ステップＳ２０９）。分析装置２１００は、繰り返しの計算回数がカウント値未満の場合には（ステップＳ２０９０，Ｙｅｓ）、ステップＳ２０４に移行する。

【０１０１】
一方、分析装置２００は、繰り返しの計算回数がカウント値以上である場合には（ステップＳ２０９，Ｎｏ）、図１４のステップＳ２１０に移行する。
【０１０２】
図１４の説明に移行する。分析装置２００は、パレート解に対応する予測データを選択して、中間データテーブル２４５を作成する（ステップＳ２１０）。分析装置２００は、パレート解に対応する各予測データの類似度を算出する（ステップＳ２１１）。分析装置２００は、類似する各予測データを、グループに分類する（ステップＳ２１２）。
【０１０３】
分析装置２００は、未選択のグループを選択し（ステップＳ２１３）、ランダムにクラスタ番号を割り振った複数の最終データ候補を生成する（ステップＳ２１４）。ステップＳ２１４において、分析装置２００は、類似度が大きくなるようにクラスタ番号を割り振る。例えば、分析装置２００は、ランダムにクラスタ番号を割り振り、類似度を計算する。そして、分析装置２００は、類似度が大きい、クラスタ番号の割り振りを少し変更して、類似度が大きくなるか、試行する処理を利用者が設定した回数繰り返す。
【０１０４】
分析装置２００は、グループに含まれる予測データと各最終データ候補とを比較して、類似度が最大となる最終データ候補を判定する（ステップＳ２１５）。
【０１０５】
分析装置２００は、未選択のグループが存在するか否かを判定する（ステップＳ２１６）。分析装置２００は、未選択のグループが存在する場合には（ステップＳ２１６，Ｙｅｓ）、ステップＳ２１３に移行する。一方、分析装置２００は、未選択のグループが存在しない場合には（ステップＳ２１６，Ｎｏ）、各グループの判定結果を出力する（ステップＳ２１７）。
【０１０６】
次に、本実施例２に係る分析装置２００の効果について説明する。分析装置２００は、複数の予測データのうち、評価値のよい予測データを類似する予測データ同士でグルーピングし、グループ毎に、最終データ２４６を生成する。このため、分析装置２００によれば、クラスタ分析に要する時間を削減することができる。また、類似する予測データに応じた最終データの候補を複数得ることが出来る。
【０１０７】
次に、上記実施例に示した分析装置１００，２００と同様の機能を実現する分析プログラムを実行するコンピュータの一例について説明する。図１５は、分析プログラムを実行するコンピュータの一例を示す図である。
【０１０８】
図１５に示すように、コンピュータ３００は、各種演算処理を実行するＣＰＵ３０１と、ユーザからのデータの入力を受け付ける入力装置３０２と、ディスプレイ３０３を有する。また、コンピュータ３００は、記憶媒体からプログラム等を読取る読み取り装置３０４と、ネットワークを介して他のコンピュータとの間でデータの授受を行うインターフェース装置３０５とを有する。また、コンピュータ３００は、各種情報を一時記憶するＲＡＭ３０６と、ハードディスク装置３０７を有する。そして、各装置３０１～３０７は、バス３０８に接続される。
【０１０９】
ハードディスク装置３０７は、サンプリングプログラム３０７ａ、クラスタ分析プログラム３０７ｂ、クラスタ予測プログラム３０７ｃ、判定プログラム３０７ｄ、最終クラスタ計算プログラム３０７ｅを有する。ＣＰＵ３０１は、各プログラム３０７ａ～３０７ｅを読み出してＲＡＭ３０６に展開する。
【０１１０】
サンプリングプログラム３０７ａは、サンプリングプロセス３０６ａとして機能する。クラスタ分析プログラム３０７ｂは、クラスタ分析プロセス３０６ｂとして機能する。ク

ラスタ予測プログラム３０７ｃは、クラスタ予測プロセス３０６ｃとして機能する。判定プログラム３０７ｄは、判定プロセス３０６ｄとして機能する。最終クラスタ計算プログラム３０７ｅは、最終クラスタ計算プロセス３０６ｅとして機能する。

【０１１１】

例えば、サンプリングプロセス３０６ａは、サンプリング実行部１５１，２５１に対応する。クラスタ分析プロセス３０６ｂは、クラスタ分析部１５２，２５２に対応する。クラスタ予測プロセス３０６ｃは、クラスタ予測部１５３，２５３に対応する。判定プロセス３０６ｄは、判定部１５４，２５４に対応する。最終クラスタ計算プロセス３０６ｅは、最終クラスタ計算部１５５，２５５に対応する。

【０１１２】

なお、各プログラム３０７ａ～３０７ｅについては、必ずしも最初からハードディスク装置３０７に記憶させておかなくても良い。例えば、コンピュータ３００に挿入されるフレキシブルディスク（ＦＤ）、ＣＤ－ＲＯＭ、ＤＶＤディスク、光磁気ディスク、ＩＣカードなどの「可搬用の物理媒体」に各プログラムを記憶させておく。そして、コンピュータ５００がこれらから各プログラム３０７ａ～３０７ｅを読み出して実行するようにしてもよい。

【０１１３】

以上の各実施例を含む実施形態に関し、さらに以下の付記を開示する。

【０１１４】

（付記１）入力データに対してサンプリングを実行し前記入力データから一部のデータを抽出する処理を繰り返し実行して複数のサンプリングデータを生成するサンプリング実行部と、

前記複数のサンプリングデータについてクラスタ分析を実行し、前記サンプリングデータ毎に、前記サンプリングデータに含まれるデータを異なるクラスタに分類するクラスタ分析部と、

前記複数のサンプリングデータに対する前記クラスタ分析部の複数の分類結果と前記入力データとを基にして、前記入力データに含まれるデータの所属するクラスタを予測したデータを示す予測データを複数生成するクラスタ予測部と、

前記予測データのクラスタ間距離およびクラスタ内距離を基にして、予測データ毎に評価値を算出し、パレート解となる評価値に対応する予測データを判定する判定部と、

前記パレート解となる評価値に対応する予測データを基にして、前記入力データに含まれるデータをクラスタに分類する最終クラスタ計算部と

を有することを特徴とする分析装置。

【０１１５】

（付記２）前記最終クラスタ計算部は、パレート解となる評価値に対応する類似の予測データをグループ化し、同一グループに含まれる予測データを基にして、前記入力データに含まれるデータを異なるクラスタに分類する処理を、グループ毎に実行することを特徴とする付記１に記載の分析装置。

【０１１６】

（付記３）前記最終クラスタ計算部は、前記入力データに対して、ランダムにクラスタを割り当てた複数の最終クラスタデータを生成し、各最終クラスタデータと予測データとの類似度を基にして、特定の最終クラスタデータを選択することを特徴とする付記１または２に記載の分析装置。

【０１１７】

（付記４）コンピュータが実行する分析方法であって、

入力データに対してサンプリングを実行し前記入力データから一部のデータを抽出する処理を繰り返し実行して複数のサンプリングデータを生成し、

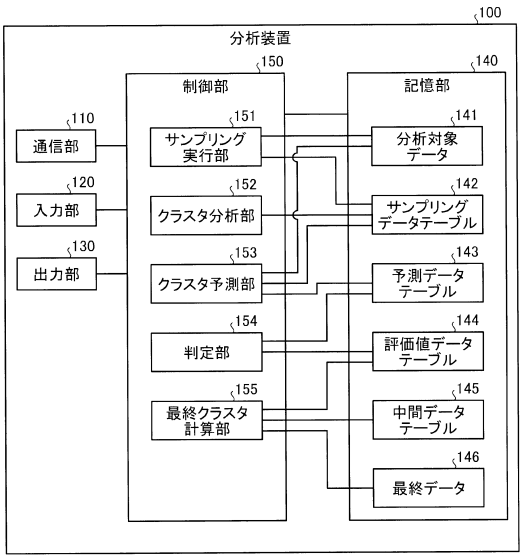
前記複数のサンプリングデータについてクラスタ分析を実行し、前記サンプリングデータ毎に、前記サンプリングデータに含まれるデータを異なるクラスタに分類し、

前記複数のサンプリングデータに対する前記クラスタ分析部の複数の分類結果と前記入

力データとを基にして、前記入力データに含まれるデータの所属するクラスタを予測したデータを示す予測データを複数生成し、
前記予測データのクラスタ間距離およびクラスタ内距離を基にして、予測データ毎に評価値を算出し、パレート解となる評価値に対応する予測データを判定し、
前記パレート解となる評価値に対応する予測データを基にして、前記入力データに含まれるデータをクラスタに分類する
各処理を実行することを特徴とする分析方法。
【０１１８】
（付記５）前記入力データに含まれるデータをクラスタに分類する処理は、パレート解となる評価値に対応する類似の予測データをグループ化し、同一グループに含まれる予測データを基にして、前記入力データに含まれるデータを異なるクラスタに分類する処理を、グループ毎に実行することを特徴とする付記４に記載の分析方法。
【０１１９】
（付記６）前記入力データに含まれるデータをクラスタに分類する処理は、前記入力データに対して、ランダムにクラスタを割り当てた複数の最終クラスタデータを生成し、各最終クラスタデータと予測データとの類似度を基にして、特定の最終クラスタデータを選択することを特徴とする付記４または５に記載の分析方法。
【０１２０】
（付記７）コンピュータに、
入力データに対してサンプリングを実行し前記入力データから一部のデータを抽出する処理を繰り返し実行して複数のサンプリングデータを生成し、
前記複数のサンプリングデータについてクラスタ分析を実行し、前記サンプリングデータ毎に、前記サンプリングデータに含まれるデータを異なるクラスタに分類し、
前記複数のサンプリングデータに対する前記クラスタ分析部の複数の分類結果と前記入力データとを基にして、前記入力データに含まれるデータの所属するクラスタを予測したデータを示す予測データを複数生成し、
前記予測データのクラスタ間距離およびクラスタ内距離を基にして、予測データ毎に評価値を算出し、パレート解となる評価値に対応する予測データを判定し、
前記パレート解となる評価値に対応する予測データを基にして、前記入力データに含まれるデータをクラスタに分類する
各処理を実行させることを特徴とする分析プログラム。
【０１２１】
（付記８）前記入力データに含まれるデータをクラスタに分類する処理は、パレート解となる評価値に対応する類似の予測データをグループ化し、同一グループに含まれる予測データを基にして、前記入力データに含まれるデータを異なるクラスタに分類する処理を、グループ毎に実行することを特徴とする付記７に記載の分析プログラム。
【０１２２】
（付記９）前記入力データに含まれるデータをクラスタに分類する処理は、前記入力データに対して、ランダムにクラスタを割り当てた複数の最終クラスタデータを生成し、各最終クラスタデータと予測データとの類似度を基にして、特定の最終クラスタデータを選択することを特徴とする付記４または５に記載の分析方法。
【符号の説明】
【０１２３】
１００，２００　分析装置
１５１，２５１　サンプリング実行部
１５２，２５２　クラスタ分析部
１５３，２５３　クラスタ予測部
１５４，２５４　判定部
１５５，２５５　最終クラスタ計算部

【図１】

本実施例1に係る分析装置の構成を示す機能ブロック図




【図２】

分析対象データのデータ構造の一例を示す図

141

| 識別番号 | 年齢 | 性別 | 身長 | 体重 | … |
|---|---|---|---|---|---|
| 1001 | 35 | 1 | 170 | 65 | … |
| 1002 | 40 | 2 | 165 | 59 | … |
| … | | | | | |

【図３】

サンプリングデータテーブルのデータ構造の一例を示す図

142

| 識別番号 | 年齢 | 性別 | 身長 | 体重 | … | クラスタ番号 |
|---|---|---|---|---|---|---|
| 1001 | 35 | 1 | 170 | 65 | … | 1 |
| 1002 | 40 | 2 | 180 | 70 | … | 2 |
| … | | | | | | |

142a 142b 142c

【図４】

予測データテーブルのデータ構造の一例を示す図

143

| 識別番号 | 年齢 | 性別 | 身長 | 体重 | … | クラスタ番号 |
|---|---|---|---|---|---|---|
| 1001 | 35 | 1 | 170 | 65 | … | 1 |
| 1002 | 40 | 2 | 165 | 59 | … | 1 |
| … | | | | | | |
| 1005 | 44 | 2 | 180 | 70 | … | 2 |
| … | | | | | | |

143a 143b 143c

【図５】

評価値データテーブルのデータ構造の一例を示す図

144

| 予測データ識別情報 | 評価値 | |
|---|---|---|
| | クラスタ間距離 | クラスタ内距離 |
| 予測データ1 | 5 | 25 |
| 予測データ2 | 10 | 20 |
| 予測データ3 | 16 | 15 |
| … | | |

【図６】

中間データテーブルのデータ構造の一例を示す図

145

| 識別番号 | クラスタ番号:1 | クラスタ番号:2 | クラスタ番号:3 |
|---|---|---|---|
| 1001 | 1 | 0 | 0 |
| 1002 | 1 | 0 | 0 |
| 1003 | 0 | 1 | 0 |
| 1004 | 0 | 0 | 1 |
| … | | | |

145a 145g 145z

【図７】

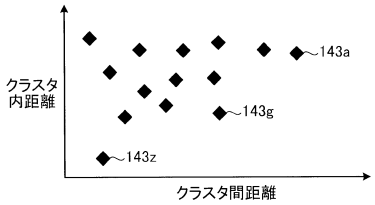
最終データのデータ構造の一例を示す図

146

| 識別番号 | クラスタ番号:1 | クラスタ番号:2 | クラスタ番号:3 |
|---|---|---|---|
| 1001 | 1 | 0 | 0 |
| 1002 | 1 | 0 | 0 |
| 1003 | 0 | 1 | 0 |
| 1004 | 0 | 0 | 1 |
| … | | | |

【図８】

各予測データのクラスタ内距離とクラスタ間距離との関係を示す図（1）




【図９】

最終データ候補テーブルの一例を示す図

10

10a

| 識別番号 | クラスタ番号:1 | クラスタ番号:2 | クラスタ番号:3 |
|---|---|---|---|
| 1001 | 1 | 0 | 0 |
| 1002 | 1 | 0 | 0 |
| 1003 | 0 | 1 | 0 |
| 1004 | 0 | 0 | 1 |
| ・・・ | | | |

10b

| 識別番号 | クラスタ番号:1 | クラスタ番号:2 | クラスタ番号:3 |
|---|---|---|---|
| 1001 | 0 | 0 | 1 |
| 1002 | 1 | 0 | 0 |
| 1003 | 0 | 1 | 0 |
| 1004 | 0 | 0 | 1 |
| ・・・ | | | |

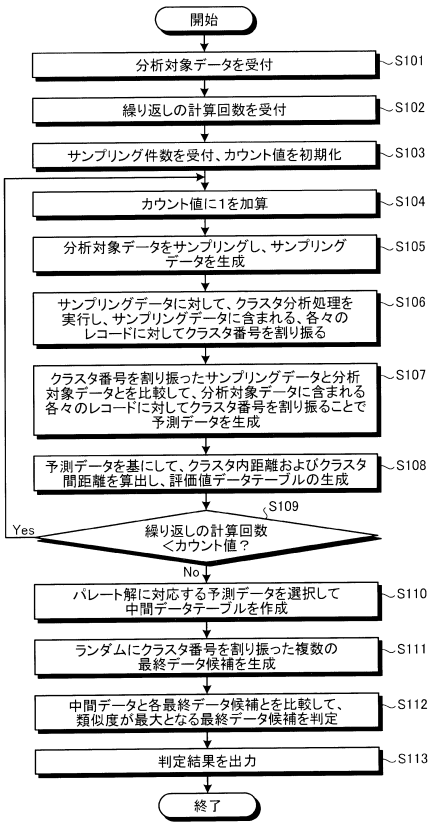
10c

| 識別番号 | クラスタ番号:1 | クラスタ番号:2 | クラスタ番号:3 |
|---|---|---|---|
| 1001 | 0 | 1 | 0 |
| 1002 | 0 | 1 | 0 |
| 1003 | 0 | 1 | 0 |
| 1004 | 0 | 0 | 1 |
| ・・・ | | | |

⋮

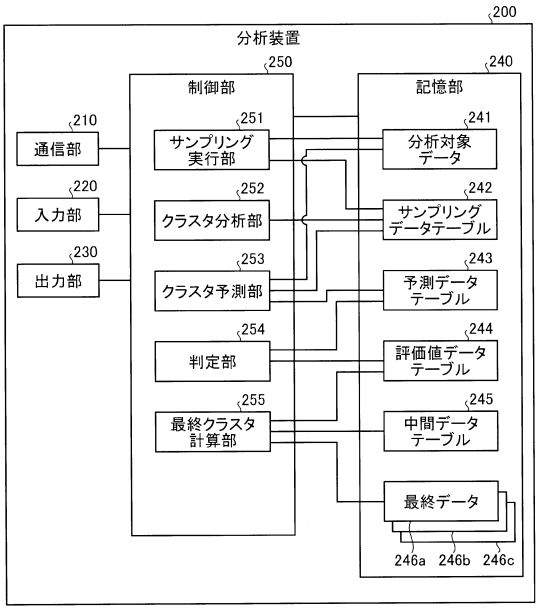
【図１０】

本実施例1にかかる分析装置の処理手順を示すフローチャート



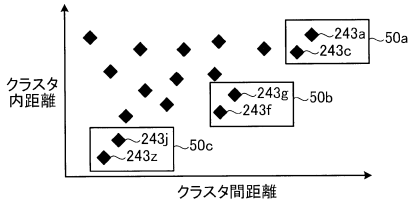

【図１１】

本実施例2にかかる分析装置の構成を示す機能ブロック図



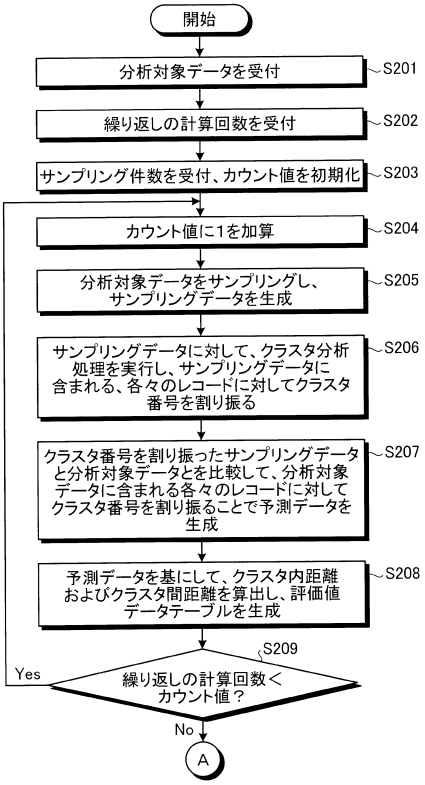
【図１２】

各予測データのクラスタ内距離とクラスタ間距離との関係を示す図（2）



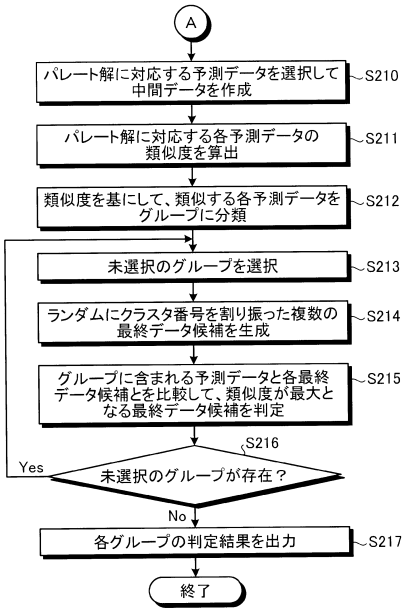

【図１３】

実施例2にかかる分析装置の処理手順を示すフローチャート（1）



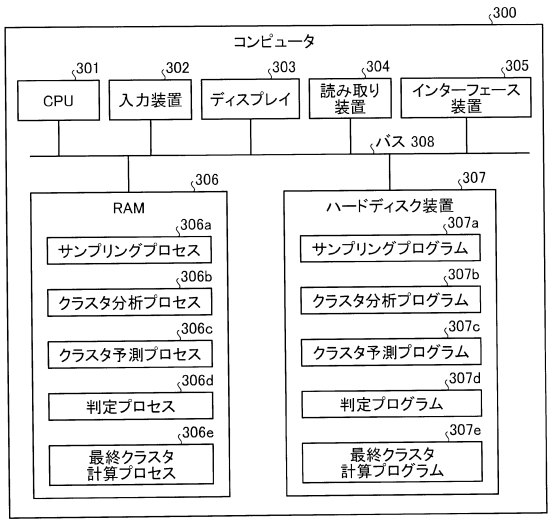

【図１４】

実施例2にかかる分析装置の処理手順を示すフローチャート（2）




【図１５】

分析プログラムを実行するコンピュータの一例を示す図



フロントページの続き

(56)参考文献 特開２００７－１７９１４３（ＪＰ，Ａ）
特開平１１－２５００９４（ＪＰ，Ａ）
米国特許第７１６２４１３（ＵＳ，Ｂ１）
特開２００７－１７９２８８（ＪＰ，Ａ）
特開２００７－１７２４２７（ＪＰ，Ａ）
小林　優、外２名，特徴要素の重みを考慮に入れたクラスタ代表の洗練による文書クラスタリング，情報処理学会研究報告，日本，社団法人情報処理学会，２００２年　3月１５日，第２００２巻，第２８号，ｐ．１３５－１４２
山代　大輔、外２名，可視化手法を用いた多目的最適化問題における満足解の選択支援，知能と情報，日本，日本知能情報ファジィ学会，２００８年１２月１５日，第２０巻，第６号，ｐ．２４－３２

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／３０