(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2019-56983
(P2019-56983A)
(43)公開日　平成31年4月11日(2019.4.11)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 16/00	(2019.01)	G06F	17/30	220Z	
G06N 20/00	(2019.01)	G06F	17/30	210D	
		G06N	99/00	150	

審査請求 未請求　請求項の数 9　ＯＬ　（全 31 頁）

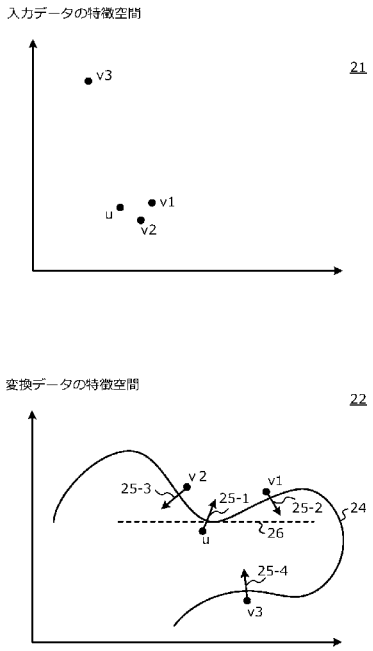
(21)出願番号　特願2017-179609 (P2017-179609)
(22)出願日　平成29年9月19日 (2017.9.19)

(71)出願人　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　110002147
特許業務法人酒井国際特許事務所
(72)発明者　後藤　啓介
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
(72)発明者　丸橋　弘治
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
(72)発明者　稲越　宏弥
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

(54)【発明の名称】学習データ選択プログラム、学習データ選択方法、および、学習データ選択装置

(57)【要約】
【課題】　変換された入力データに対する分類・判別の要因を推定する機械学習モデルの、推定精度を向上させる。
【解決手段】
入力データを変換した変換データに対し分類・判別を行う機械学習モデルの出力要因を推定する、推定モデルの学習データの選択を行うために、１）機械学習モデルへの入力データ群に含まれる第１の入力データの指定に応じ、第１の入力データに関連する第１の入力データ群を抽出し、２）第１の入力データ群に対応した、機械学習モデルに入力される第１の変換データ群、および、第１の変換データ群に対応した機械学習モデルの第１の出力データ群を、それぞれ取得し、３）第１の入力データと第１の入力データ群のデータそれぞれとの距離、および、第１の変換データと第１の変換データ群のデータそれぞれとの距離に基づき、第１の入力データ群から、推定モデルの学習対象データを選択する。

【選択図】　図３



【特許請求の範囲】
【請求項１】
入力データを変換した変換データに対し分類または判別を行う機械学習モデルの出力要因を推定する、推定モデルの学習データの選択を、コンピュータに実行させる学習データ選択プログラムであって、
前記機械学習モデルへの入力データ群に含まれる第１の入力データの指定に応じ、前記第１の入力データに関連する第１の入力データ群を抽出し、
前記第１の入力データ群に対応した、前記機械学習モデルに入力される第１の変換データ群、および、前記第１の変換データ群に対応した前記機械学習モデルの第１の出力データ群を、それぞれ取得し、
前記第１の入力データと前記第１の入力データ群のデータそれぞれとの距離、および、前記第１の変換データと前記第１の変換データ群のデータそれぞれとの距離に基づき、前記第１の入力データ群から、前記推定モデルの学習対象データを選択する、
ことを特徴とする学習データ選択プログラム。

【請求項２】
請求項１記載の学習データ選択プログラムであって、
前記抽出する処理は、前記第１の出力データ群に含まれる出力データそれぞれのデータ内容に基づき、学習対象データを抽出する、
ことを特徴とする学習データ選択プログラム。

【請求項３】
請求項２記載の学習データ選択プログラムであって、
前記抽出する処理は、前記第１の出力データ群に含まれるデータ内容の比率に基づき、学習対象データを抽出する、
ことを特徴とする学習データ選択プログラム。

【請求項４】
請求項３記載の学習データ選択プログラムであって、
前記抽出する処理は、前記第１の出力データ群に含まれるデータ内容の正例と負例の比率に基づき、学習対象データを抽出する、
ことを特徴とする学習データ選択プログラム。

【請求項５】
請求項１記載の学習データ選択プログラムであって、
前記特定する処理は、データ取得タイミングが前記第１の入力データと所定の関係を有する入力データを前記第１の入力データ群と特定する、
ことを特徴とする学習データ選択プログラム。

【請求項６】
請求項１記載の学習データ選択プログラムであって、
前記特定する処理は、データ生成元が前記第１の入力データと所定の関係を有する入力データを前記第１の入力データ群と特定する、
ことを特徴とする学習データ選択プログラム。

【請求項７】
請求項１記載の学習データ選択プログラムであって、
前記抽出する処理は、前記第１の入力データと前記第１の入力データ群のデータそれぞれとの距離を、前記第１の入力データと前記第１の入力データ群のデータそれぞれとを個別に変換して算出する、

ことを特徴とする学習データ選択プログラム。

【請求項８】
入力データを変換した変換データに対し分類または判別を行う機械学習モデルの出力要因を推定する、推定モデルの学習データの選択方法であって、
前記機械学習モデルへの入力データ群に含まれる第１の入力データの指定に応じ、前記第１の入力データに関連する第１の入力データ群を抽出し、
前記第１の入力データ群に対応した、前記機械学習モデルに入力される第１の変換データ群、および、前記第１の変換データ群に対応した前記機械学習モデルの第１の出力データ群を、それぞれ記憶装置より取得し、
前記第１の入力データと前記第１の入力データ群のデータそれぞれとの距離、および、前記第１の変換データと前記第１の変換データ群のデータそれぞれとの距離に基づき、前記第１の入力データ群から、前記推定モデルの学習対象データを選択する、
ことを特徴とする学習データ選択方法。

【請求項９】
入力データを変換した変換データに対し分類または判別を行う機械学習モデルの出力要因を推定する、推定モデルの学習データの選択装置であって、
前記機械学習モデルへの入力データ群に含まれる第１の入力データの指定に応じ、前記第１の入力データに関連する第１の入力データ群を抽出する抽出部と、
前記第１の入力データ群に対応した、前記機械学習モデルに入力される第１の変換データ群、および、前記第１の変換データ群に対応した前記機械学習モデルの第１の出力データ群を、それぞれ記憶装置より取得する選択部と、を有し、
前記選択部は、前記第１の入力データと前記第１の入力データ群のデータそれぞれとの距離、および、前記第１の変換データと前記第１の変換データ群のデータそれぞれとの距離に基づき、前記第１の入力データ群から、前記推定モデルの学習対象データを選択する、
ことを特徴とする学習データ選択装置。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、機械学習の学習対象データを選択する、学習データ選択プログラム、学習データ選択方法、および、学習データ選択装置に関する。
【背景技術】
【０００２】
近年、さまざまな分野のデータに対して、機械学習を用いた分類や判別が行われている。機械学習を用いることにより、精度の高い分類や判別が可能である一方、機械学習のどの特徴量が分類や判別の要因であるかは、一般的には知ることができない。
【０００３】
機械学習が分類や判別をする際に、どのような要因により分類や判別を行ったのかが明確でない場合、たとえば、機械学習の適用分野を拡大する際のリスクとなる。
【０００４】
機械学習に用いられる分類器による分類結果から、分類の要因となる特徴を推定する方法として、ＬＩＭＥ（Local Interpretable Model-agnostic Explanations）という手法が知られている（例えば非特許文献１参照）。ＬＩＭＥにおいては、任意の分類器ｆと入力データｕについて、ｕの分類結果ｆ（ｕ）に大きく貢献したｕの要素・特徴を推定し、提示することが行われる。
【０００５】
非特許文献１に記載されたＬＩＭＥの手法では、画像を対象とした分類結果に対する原

因推定が行われており、どの画像の部分が分類結果に寄与したかを推定することが記載されている。

【０００６】

図２は、分類器による入力データの分類と、分類要因の推定の概要を示す図である。一般的な機械学習の分類器を用いた分類では、入力データを分類器が処理可能な次元に変換した変換データを生成し、生成された変換データに対して学習および分類が行われる。

【０００７】

図２に示される一般的な機械学習の分類器を用いた分類に対し、ＬＩＭＥの手法により分類結果に対する分類要因の推定を行うと、分類結果が変換データのどの要因に関連するかを推定するものとなり、入力データのどの要因に関連するものかを推定するものではないため、必ずしも有用であるとは限らなかった。

【先行技術文献】

【非特許文献】

【０００８】

【非特許文献１】“Why Should I Trust You?” Explaining the Predictions of Any Classifier、Marco Tulio Ribeiro et. al.、Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining、2016年8月

【発明の概要】

【発明が解決しようとする課題】

【０００９】

上述したように、入力データを分類器が処理可能な次元に変換した変換データを生成し、生成された変換データに対して学習および分類が行われる機械学習の分類器では、入力データにおける分類要因の推定が求められる。

【００１０】

図３は、入力データから生成された変換データの分類を行う機械学習の分類器に関する、各特徴空間における分類器への入力データと変換データの概要を示す図である。図３において、点ｕは正例として判別されるデータ、点ｖ１、点ｖ２、点ｖ３は負例として判別されるデータに対応する、入力データの特徴空間２１および変換データの特徴空間２２上での点である。

【００１１】

特徴空間２２において、点ｕ、点ｖ１、点ｖ２、点ｖ３にはそれぞれ、予測値が大きく変動する方向を表すベクトルである説明ベクトル２５－１，２５－２，２５－３、２５－４がそれぞれ示されている。また、変換データの特徴空間２２において、近似識別線２４は正例と負例の識別境界を近似する境界線である。入力データの特徴空間２１では必ずしも分類器の正例と負例の識別境界が明確とは限らないため、図３の入力データの特徴空間２１では近似識別線は描かれてはいない。

【００１２】

ここで、図２に示した分類要因の推定を図３の点ｕの近傍について行うことは、正例と負例とを識別する近似識別線２４の点ｕの近傍における識別要因を示す識別要因２６を推定することに相当する。

【００１３】

データｕの近傍における分類要因を推定する場合、入力データの特徴空間２１では、点ｕ、点ｖ１、点ｖ２と点ｖ３とは離れているため、点ｖ３をＬＩＭＥの手法における点ｕの近傍の点として扱うことは適切ではない。

【００１４】

非特許文献１に記載されたＬＩＭＥの手法により分類要因を推定する場合、変換データの特徴空間２２内において点ｕの近傍の点を選択し、分類要因を推定する。具体的には、変換データの特徴空間２２で点ｕの近傍に存在する点ｖ１、点ｖ２、点ｖ３それぞれを学習データとして機械学習した推定器により、分類要因の推定を行う。すなわち、点ｖ３という、変換データの特徴空間２２においては点ｕの近傍にあるものの、入力データの特徴

空間２１においては点ｕの近傍にあるとはいえない点を学習データに含めて学習が行われるため、学習された推定モデルによる判別精度が劣化するという問題がある。

【００１５】

１つの側面では、本件は、変換された入力データに対する分類または判別の要因を推定する機械学習モデルの、推定精度を向上させることを目的とする。

【課題を解決するための手段】

【００１６】

１つの案では、入力データを変換した変換データに対し分類または判別を行う機械学習モデルの出力要因を推定する、推定モデルの学習データの選択について、コンピュータに以下の処理を実行させる学習データ選択プログラム、学習データ選択方法、および、学習データ選択装置が提供される。

【００１７】

すなわち、コンピュータに、機械学習モデルへの入力データ群に含まれる第１の入力データの指定に応じ、第１の入力データに関連する第１の入力データ群を抽出し、
第１の入力データ群に対応した、機械学習モデルに入力される第１の変換データ群、および、第１の変換データ群に対応した機械学習モデルの第１の出力データ群を、それぞれ取得し、
第１の入力データと第１の入力データ群のデータそれぞれとの距離、および、第１の変換データと第１の変換データ群のデータそれぞれとの距離に基づき、第１の入力データ群から、推定モデルの学習対象データを選択することを特徴とする。

【発明の効果】

【００１８】

１態様によれば、出力要因の推定精度を向上させる学習データを選択することができる。

【図面の簡単な説明】

【００１９】

【図１】第１の実施形態に係る分類装置および要因推定装置の構成例を示す図である。
【図２】分類器による入力データの分類と、分類要因の推定の概要を示す図である。
【図３】各特徴空間における分類器への入力データと変換データの概要を示す図である。
【図４】第１の実施形態に係る処理の手順の一例を示すフローチャートである。
【図５】第１の実施形態のシステム構成例を示す図である。
【図６】第１の実施形態に用いる監視サーバのハードウェアの一構成例を示す図である。
【図７】第２の実施形態に係る要因推定装置の構成例を示す図である。
【図８Ａ】選択データｕについての入力データと中間データとの関係を示す図である。
【図８Ｂ】対象データｖ１～ｖ３についての入力データと中間データとの関係を示す図である。
【図９Ａ】選択データｕについての入力データと変換データとの関係を示す図である。
【図９Ｂ】対象データｖ１～ｖ３についての入力データと変換データとの関係を示す図である。
【図１０】選択データｕと対象データｖ１～ｖ３の各特徴空間における距離を示す図である。
【図１１】第２の実施形態における入力データ記憶部の一例を示す図である。
【図１２】第２の実施形態における変換データ記憶部の一例を示す図である。
【図１３】第２の実施形態における解析結果記憶部の一例を示す図である。
【図１４】第２の実施形態における入力データ距離記憶部の一例を示す図である。
【図１５】第２の実施形態における変換データ距離記憶部の一例を示す図である。
【図１６】第２の実施形態に係る処理の手順の一例を示すフローチャートである。
【図１７】第２の実施形態に係る入力データ距離算出部２２１の距離算出の一例を示す図である。
【図１８】変換行列を用いて生成した中間データ間の類似度の計算例を示す図である。

【図１９】「項Ｓ」の変換行列の更新例を示す図である。
【図２０】類似度計算処理の手順の一例を示すフローチャートである。
【図２１】初期状態の変換行列を用いた中間データの生成例を示す図である。
【図２２】「項Ｓ」の変換行列の更新例を示す図である。
【図２３】「項Ｒ」の変換行列の更新例を示す図である。
【図２４】更新後の変換行列を用いた中間データの生成例を示す図である。
【図２５】類似度計算の比較例を示す第１の図である。
【図２６】類似度計算の比較例を示す第２の図である。
【図２７】類似度計算例を示す第１の図である。
【図２８】類似度計算例を示す第２の図である。
【発明を実施するための形態】
【００２０】
以下、本実施の形態について図面を参照して説明する。なお各実施の形態は、矛盾のない範囲で複数の実施の形態を組み合わせて実施することができる。
【００２１】
〔第１の実施形態〕
図１は、第１の実施形態に係る分類装置１０および要因推定装置２０の構成例を示す図である。分類装置１０は、収集部１１０、入力データ記憶部１１２、変換部１２０、変換データ記憶部１２２、分類部１３０、学習結果記憶部１３２、解析部１４０、解析結果記憶部１４２を有する。
【００２２】
要因推定装置２０は、抽出部２１０、選別部２２０、選別データ記憶部２３０、推定部２４０、学習結果記憶部２４２を有する。例えばコンピュータが、学習データ選別プログラムを実行することによって、学習データの選別方法を実行可能な要因推定装置２０が実現される。
【００２３】
分類装置１０において、収集部１１０は、入力データを収集し、入力データ記憶部１１２に記憶する。たとえば、図３に示される入力データの特徴空間２１の入力データｕ、ｖ１、ｖ２、ｖ３を含む入力データが入力データ記憶部１１２に記憶される。変換部１２０は、収集部１１０で収集された入力データを所定の方法により変換し、変換データ記憶部１２２に記憶する。たとえば、図３に示される入力データの特徴空間２１の入力データｕ、ｖ１、ｖ２、ｖ３は、変換データの特徴空間２２のｕ、ｖ１、ｖ２、ｖ３にそれぞれ変換される。所定の方法とは、たとえば、分類部１３０における分類が適切に行われるように入力データの次元、値等を変換するものであり、線形変換、非線形変換、一方向変換、双方向変換などを用いることができる。分類部１３０は、変換部１２０で変換された変換データを入力し、学習結果記憶部１３２に記憶された学習パラメータを用いて分類を行い、解析部１４０に出力する。
【００２４】
要因推定装置２０において、抽出部２１０は、たとえばユーザから、要因推定対象となる入力データの指定を受け、分類装置１０の入力データ記憶部１１２から関連するデータを取得する。選別部２２０は、抽出部２１０で取得されたデータについて、分類装置１０の変換データ記憶部１２２から変換データを、分類装置１０の解析結果記憶部１４２から変換データに対応した解析結果を取得し、推定部２４０の学習対象となる学習データを選別する。推定部２４０は、選別部２２０により選別された学習データに基づき学習を行い、学習結果記憶部２４２に学習パラメータを記憶する。
【００２５】
図４は、第１の実施形態に係る処理の手順の一例を示すフローチャートである。図１に示された要因推定装置２０の抽出部２１０は、たとえばユーザから要因推定対象となる入力データである選択データとして、図３に示される点ｕに対応した入力データｕの指定を受ける。入力データｕの指定を受けたことに対応して、抽出部２１０は、所定の基準によ

り分類装置１０の入力データ記憶部１１２より、図３に示された点ｖ１～点ｖ３に対応した対象データである入力データｖ１～ｖ３を抽出する（Ｓ１２）。所定の基準としては、例えば、入力データの収集タイミングが選択データｕと所定の関係を有するもの、具体的には、データ収集期間が選択データｕと前後１時間のもの、とすることができる。また、所定の基準として、例えば、データ生成元が選択データｕと所定の関係を有するもの、具体的には、同じサーバから対象データが取得されたもの、とすることもできる。

【００２６】

次に、抽出部２１０からの選択データおよび対象データの抽出を受け、選別部２２０は、選択データｕおよび対象データｖ１～ｖ３に対応する変換データおよび分類結果を、分類装置１０の変換データ記憶部１２２および解析結果記憶部１４２より読み込む（Ｓ１４）。

【００２７】

次に、選別部２２０は、選択データｕと対象データｖ１～ｖ３それぞれとの、入力データの特徴空間における距離と、変換データの特徴空間における距離とを算出し、距離に基づいて学習対象のデータを選別する（Ｓ１６）。

【００２８】

図３に示された選択データｕ、および、対象データｖ１～ｖ３は、入力データの特徴空間２１においては、点ｕと点ｖ１、点ｖ２との距離に比べ、点ｕと点ｖ３との距離は大きい。一方、変換データの特徴空間２２においては、点ｕと点ｖ１～点ｖ３との距離は大きくは変わらない。

【００２９】

すなわち、変換データの特徴空間２２では点ｕと近傍にある点ｖ１～点ｖ３のうち、点ｖ３は入力データの特徴空間２１では点ｕの近傍とはならない。このため、対象データｖ３は、分類要因を推定する推定部２４０の学習データとしては不適切である。

【００３０】

入力データの特徴空間２１における点ｕに対する距離と、変換データの特徴空間２２における点ｕに対する距離の和を取ると、対象データｖ３の距離の和は、対象データｖ１の距離の和、または、選択データｖ２の距離の和よりも大きくなるから、距離の和の計算結果に基づき、対象データｖ３を除外し、対象データｖ１およびｖ２を学習対象として選択することができる。

【００３１】

次に、選別されたデータを用いて、推定部２４０において学習が行われる（Ｓ１８）。すなわち、推定部２４０において、選択データｕと対象データｖ１およびｖ２の入力データ、変換データ、および、分類結果に基づく学習が行われ、学習結果に対応する学習パラメータが、学習結果記憶部２４２に記憶される。

【００３２】

図５は、第１の実施形態のシステム構成例を示す図である。ネットワーク２０００には、複数のサーバ２０１１，２０１２，・・・、複数の端末装置２０２１，２０２２，・・・、および監視サーバ１００１が接続されている。複数のサーバ２０１１，２０１２，・・・は、いずれかの端末装置からの要求に応じた処理を実行するコンピュータである。複数のサーバ２０１１，２０１２，・・・のうちの２台以上が連携して処理を実行する場合もある。複数の端末装置２０２１，２０２２，・・・は、複数のサーバ２０１１，２０１２，・・・で提供されるサービスを利用するユーザが使用するコンピュータである。

【００３３】

監視サーバ１００１は、ネットワーク２０００を介して行われた通信を監視し、通信ログを記録する。監視サーバ１００１は、単位時間帯ごとの通信ログのデータを分類する。例えば監視サーバ１００１は、対応する単位時間帯における不正通信の有無に応じて、データを分類する。

【００３４】

監視サーバ１００１上で動作する分類装置１０の収集部１１０は、ネットワーク２００

０を介して送受信されているパケットなどの通信情報を取得する。例えば収集部１１０は、ネットワーク２０００内に設置されたスイッチのミラーリングポートを介して、そのスイッチを経由して通信されたパケットを取得する。また収集部１１０は、各サーバ２０１１，２０１２，・・・から、そのサーバ自身の通信ログを取得することもできる。収集部１１０は、取得した通信情報のログ（通信ログ）を、入力データ記憶部１１２に格納する。

【００３５】

なお、図５に示した各要素間を接続する線は通信経路の一部を示すものであり、図示した通信経路以外の通信経路も設定可能である。また、図５に示した各要素の機能は、例えば、その要素に対応するプログラムモジュールをコンピュータに実行させることで実現することができる。

【００３６】

図６は、第１の実施形態に用いる監視サーバのハードウェアの一構成例を示す図である。監視サーバ１０１０は、プロセッサ１０１１によって装置全体が制御されている。プロセッサ１０１１には、バス１０１９を介してメモリ１０１２と複数の周辺機器が接続されている。プロセッサ１０１１は、マルチプロセッサであってもよい。プロセッサ１０１１は、例えばＣＰＵ（Central Processing Unit）、ＭＰＵ（Micro Processing Unit）、またはＤＳＰ（Digital Signal Processor）である。プロセッサ１０１１がプログラムを実行することで実現する機能の少なくとも一部を、ＡＳＩＣ（Application Specific Integrated Circuit）、ＰＬＤ（Programmable Logic Device）などの電子回路で実現してもよい。

【００３７】

メモリ１０１２は、監視サーバ１０１０の主記憶装置として使用される。メモリ１０１２には、プロセッサ１０１１に実行させるＯＳ（Operating System）のプログラムやアプリケーションプログラムの少なくとも一部が一時的に格納される。また、メモリ１０１２には、プロセッサ１０１１による処理に必要な各種データが格納される。メモリ１０１２としては、例えばＲＡＭ（Random Access Memory）などの揮発性の半導体記憶装置が使用される。

【００３８】

バス１０１９に接続されている周辺機器としては、ストレージ装置１０１３、グラフィック処理装置１０１４、入力インタフェース１０１５、光学ドライブ装置１０１６、機器接続インタフェース１０１７およびネットワークインタフェース１０１８がある。

【００３９】

ストレージ装置１０１３は、内蔵した記録媒体に対して、電気的または磁気的にデータの書き込みおよび読み出しを行う。ストレージ装置１０１３は、コンピュータの補助記憶装置として使用される。ストレージ装置１０１３には、ＯＳのプログラム、アプリケーションプログラム、および各種データが格納される。なお、ストレージ装置１０１３としては、例えばＨＤＤ（Hard Disk Drive）やＳＳＤ（Solid State Drive）を使用することができる。

【００４０】

グラフィック処理装置１０１４には、モニタ１０２１が接続されている。グラフィック処理装置１０１４は、プロセッサ１０１１からの命令に従って、画像をモニタ１０２１の画面に表示させる。モニタ１０２１としては、ＣＲＴ（Cathode Ray Tube）を用いた表示装置や液晶表示装置などがある。

【００４１】

入力インタフェース１０１５には、キーボード１０２２とマウス１０２３とが接続されている。入力インタフェース１０１５は、キーボード１０２２やマウス１０２３から送られてくる信号をプロセッサ１０１１に送信する。なお、マウス１０２３は、ポインティングデバイスの一例であり、他のポインティングデバイスを使用することもできる。他のポインティングデバイスとしては、タッチパネル、タブレット、タッチパッド、トラックボ

ールなどがある。
【００４２】
光学ドライブ装置１０１６は、レーザ光などを利用して、光ディスク１０２４に記録されたデータの読み取りを行う。光ディスク１０２４は、光の反射によって読み取り可能なようにデータが記録された可搬型の記録媒体である。光ディスク１０２４には、ＤＶＤ（Digital Versatile Disc）、ＤＶＤ－ＲＡＭ、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）、ＣＤ－Ｒ（Recordable）／ＲＷ（ReWritable）などがある。
【００４３】
機器接続インタフェース１０１７は、監視サーバ１０１０に周辺機器を接続するための通信インタフェースである。例えば機器接続インタフェース１０１７には、メモリ装置１０２５やメモリリーダライタ１０２６を接続することができる。メモリ装置１０２５は、機器接続インタフェース１０１７との通信機能を搭載した記録媒体である。メモリリーダライタ１０２６は、メモリカード１０２７へのデータの書き込み、またはメモリカード１０２７からのデータの読み出しを行う装置である。メモリカード１０２７は、カード型の記録媒体である。
【００４４】
ネットワークインタフェース１０１８は、ネットワーク１０２０に接続されている。ネットワークインタフェース１０１８は、ネットワーク１０２０を介して、他のコンピュータまたは通信機器との間でデータの送受信を行う。
【００４５】
以上のようなハードウェア構成によって、第１の実施形態の処理機能を実現することができる。
【００４６】
監視サーバ１０１０は、例えばコンピュータが読み取り可能な記録媒体に記録されたプログラムを実行することにより、第２の実施形態の処理機能を実現する。監視サーバ１０１０に実行させる処理内容を記述したプログラムは、様々な記録媒体に記録しておくことができる。例えば、監視サーバ１０１０に実行させるプログラムをストレージ装置１０１３に格納しておくことができる。プロセッサ１０１１は、ストレージ装置１０１３内のプログラムの少なくとも一部をメモリ１０１２にロードし、プログラムを実行する。また監視サーバ１０１０に実行させるプログラムを、光ディスク１０２４、メモリ装置１０２５、メモリカード１０２７などの可搬型記録媒体に記録しておくこともできる。可搬型記録媒体に格納されたプログラムは、例えばプロセッサ１０１１からの制御により、ストレージ装置１０１３にインストールされた後、実行可能となる。またプロセッサ１０１１が、可搬型記録媒体から直接プログラムを読み出して実行することもできる。
【００４７】
以上のようなハードウェア構成によって、第１の実施形態の処理機能を実現することができる。
【００４８】
〔第２の実施形態〕
次に第２の実施形態について説明する。
【００４９】
図７は、第２の実施形態に係る要因推定装置の構成例を示す図である。図７に示された第２の実施形態に係る要因推定装置２０のうち、第１の実施形態に係る要因推定装置１０と同じ動作の部分については説明を省略する。
【００５０】
第２の実施形態に係る要因推定装置２０の選別部２２０は、入力データ距離算出部２２１、入力データ距離記憶部２２２、変換データ距離算出部２２３、変換データ距離記憶部２２４、対象判定部２２５を有する。
【００５１】
図１１は、第２の実施形態における、入力データ記憶部の一例を示す図である。図１に

示された分類装置１０に含まれる入力データ記憶部１１２には、複数の単位期間ログ１１２－１，１１２－２，・・・が格納されている。単位期間ログ１１２－１，１１２－２，・・・それぞれには、通信ログの収集期間が示されており、例えば、単位期間ログ１１２－１の収集期間は、１０：００－１０：１０、単位期間ログ１１２－２の収集期間は１０：１０－１０：２０である。単位期間ログ１１２－１，１１２－２，・・・には、収集期間で示された時間帯内に収集した通信情報が格納される。

【００５２】

単位期間ログ１１２－１，１１２－２，・・・に格納される各レコードには、通信元ホスト、通信先ホスト、ポート、および量が含まれる。通信元ホストは、パケットの送信元の装置の識別子である。通信先ホストは、パケットの宛先の装置の識別子である。ポートは、通信元ホストと通信先ホストが通信を行った通信ポート番号の識別子である。単位期間ログ１１２－１における量は、通信元ホスト・通信先ホスト・ポートの組み合わせに対する値であり、たとえば、通信元ホスト・通信先ホスト・ポートの組が同じ通信の出現回数である。

【００５３】

図１２は、第２の実施形態における、変換データ記憶部の一例を示す図である。図１に示された分類装置１０に含まれる変換データ記憶部１２２は、変換データテーブル１２２－１，１２２－２，・・・を記憶している。各変換テーブル１２２－１，１２２－２、・・・は、入力データ記憶部１１２に記憶された単位期間ログ１１２－１，１１２－２，・・・を、分類部１３０における学習および分類に対応した変換により変換したデータである。

【００５４】

入力データ記憶部１１２に記憶された単位期間ログ１１２－１，１１２－２，・・・は、図１２の各変換データの下部に示された順序づけ１２７－１，１２７－２，・・・により、変換データテーブル１２２－１，１２２－２，・・・に変換されている。

【００５５】

分類装置１０の学習段階において、分類部１３０の学習パラメータと各順序付けは、学習結果に応じた値および関係となっている。

【００５６】

図１３は、第２の実施形態における、解析結果記憶部の一例を示す図である。図１に示された分類装置１０に含まれる解析結果記憶部１４２は、入力データ記憶部１１２に記憶された、収集期間の異なる単位期間ログ１１２－１，１１２－２，・・・ごとの、分類部１３０による分類結果に対応した解析結果を記憶している。例えば、図１１の入力データ記憶部１１２に記憶された、収集期間が１０：００－１０：１０である単位期間ログ１１２－１は、分類部１３０による分類結果に基づく解析結果が、１（問題あり）として、図１３に示される解析結果テーブル１４２－１に保存されている。

【００５７】

図１６は、第２の実施形態に係る処理の手順の一例を示すフローチャートである。以下、第２の実施形態における、要因推定装置２０の選別部２２０による、学習対象データの選別手順を、図１６に示されたフローチャートと、図８Ａ、図８Ｂ、図９Ａ、図９Ｂ、および、図１０を用いて説明する。

【００５８】

図７に示された要因推定装置２０の抽出部２１０は、たとえばユーザから要因推定対象となる入力データである選択データｕとして、図１１に示された収集期間が１０：００－１０：１０である単位期間ログ１１２－１の指定を受ける。

【００５９】

選択データｕの指定に対応して、抽出部２１０は、収集期間の始期が選択データｕに続く１０：１０－１０：３０である、図１１に示された単位期間ログ１１２－２～１１２－４を対象データｖ１～ｖ３として選択する（Ｓ２２）。

【００６０】

ここでは、対象データの選定方法として、データ収集期間が選択データｕと連続するものを選択しているが、実施例１と同様に、データ収集期間に関する他の基準や、データ生成元が選択データｕと所定の関係を有するものを対象データとして選択してもよい。具体的には、同じサーバから対象データが取得されたものを対象データとして選択してもよい。

【００６１】

次に、抽出部２１０により抽出された選択データｕおよび対象データｖ１～ｖ３について、入力データ距離算出部２２１は、選択データおよび各対象データを変換する順序付けを生成する（Ｓ２４）。図８Ａ、および、図８Ｂは、選択データｕおよび対象データｖ１～ｖ３についての、入力データと中間データとの関係を示す図である。図８Ａおよび図８Ｂにおいて、選択データｕの入力データ８０１、および、対象データｖ１～ｖ３の入力データ８１１～８３１は、順序付け８０３～８３３により中間データ８０５～８３５に変換される。順序付け８０３～８３３の算出方法については、図１７～図２８を参照し後述する。

【００６２】

順序付け８０３～８３３を算出することにより、選択データｕの入力データ８０１、および、対象データｖ１～ｖ３の入力データ８１１～８３１は、中間データ８０５～８３５に変換され、入力データの特徴空間２１における選択データｕと対象データｖ１～ｖ３の距離を、それぞれ求めることができる。

【００６３】

たとえば、選択データｕと対象データｖ１との、入力データの特徴空間２１における距離は、中間データ８０５と中間データ８１５との距離を求めることにより算出される。

【００６４】

中間データ８０５と中間データ８１５についての計算例を以下に示す。

ａ）ｕとｖ１の双方に存在するが量が異なる項目は、（Ｓ’２、Ｒ’３、Ｐ’２）の１項目であるので、$(2-1)^2=2$。

ｂ）ｕのみに存在する項目は、（Ｓ’１、Ｒ’３、Ｐ’１）および（Ｓ’３、Ｒ’１、Ｐ’１）の２項目であり、$1^2+1^2=2$。

ｃ）ｖ１のみに存在する項目は、（Ｓ’２、Ｒ’３、Ｐ’１）および（Ｓ’２、Ｒ’２、Ｐ’１）の２項目であり、$1^2+1^2=2$。　となるから、ａ）ｂ）ｃ）の合計は５となる。

【００６５】

同様に計算することにより、入力データの特徴空間２１における選択データｕと対象データｖ１～ｖ３の距離は、それぞれ、５、４、９となる。

【００６６】

次に、抽出部２１０からの選択データおよび対象データの抽出を受け、選別部２２０は、選択データｕおよび対象データｖ１～ｖ３に対応する変換データおよび分類結果を、図１に示された分類装置１０の変換データ記憶部１２２および解析結果記憶部１４２より読み込む（Ｓ２６）。

【００６７】

次に、入力データの特徴空間２１におけるテンソル間距離および変換データの特徴空間２２におけるテンソル間距離に基づき学習対象のデータを選別する（Ｓ２８）。図９Ａ、および、図９Ｂは、選択データｕおよび対象データｖ１～ｖ３についての、入力データと変換データとの関係を示す図である。図９Ａおよび図９Ｂにおいて、選択データｕの入力データ９０１、および、対象データｖ１～ｖ３の入力データ９１１～９３１は、分類装置１０の変換部１２０により生成された順序付け９０３～９３３により、変換データ９０５～９３５に変換されている。

【００６８】

入力データの特徴空間２１における距離算出と同様に、変換データの特徴空間２２における選択データｕと対象データｖ１～ｖ３の距離を求める。選択データｕの変換データ９

０５と、対象データｖ１～ｖ３の変換データ９１５～９３５との距離は、それぞれ、９、８、９となる。

【００６９】

図１４は、第２の実施形態における、入力データ距離記憶部２２２の一例を示す図である。入力データ距離記憶部２２２には、入力データ距離算出部により算出された距離が、入力データ距離テーブル２２２－１として、収集期間毎に記憶される。

【００７０】

図１５は、第２の実施形態における、変換データ距離記憶部２２４の一例を示す図である。変換データ距離記憶部２２４には、変換データ距離算出部により算出された距離が、変換データ距離テーブル２２４－１として、収集期間毎に記憶される。

【００７１】

図１０は、選択データｕと対象データｖ１～ｖ３の各特徴空間における距離を示す図である。図１０において、選択データｕに対する、対象データｖ１～ｖ３の各特徴空間における距離の和は、それぞれ、１４、１２、１８となる。

【００７２】

ここで、図１０および図３を参照すると、選択データｕ、および、対象データｖ１～ｖ３は、入力データの特徴空間２１においては、点ｕと点ｖ１、点ｖ２との距離（４）に比べ、点ｕと点ｖ３との距離（９）は大きい。一方、変換データの特徴空間２２においては、点ｕと点ｖ１～点ｖ３との距離は大きくは変わらない（９または８）。

【００７３】

対象データｖ１～ｖ３について、選択データｕに対する各特徴空間の距離の和を取ると、上述したように、１４、１２、１８となるから、対象データｖ３は、対象データｖ１およびｖ２と比較して、選択データｕと離れており、分類要因を推定する推定部２４０の学習データとしては不適切である。距離の和の計算結果に基づき、対象データｖ３を除外し、選択データｖ１およびｖ２を学習対象として選択することができる。

【００７４】

図１７は、第２の実施形態に係る入力データ距離算出部２２１の距離算出の一例を示す図である。入力データ距離算出部２２１は、抽出部２１０からの入力データについて算出した類似度に基づき、距離を算出する。

【００７５】

以下では、簡単のために類似度を計算する第１データ１と第２データ２が、それぞれ２つの項目を持つ場合について説明する。項目数が３以上の場合であっても、以下の説明における行列をテンソルに拡張することで類似度を計算することが可能である。

【００７６】

図１７において、第１データ１と第２データ２は、入力データ距離算出部２２１による類似度の算出対象である。第１データ１は複数の第１レコードを有し、第１レコードの各々は、第１項目である「項Ｓ，項Ｒ」のそれぞれについての第１項目値（「項Ｓ」についての「Ｓ１，Ｓ２」、「項Ｒ」についての「Ｒ１，Ｒ２」と、第１項目値間の関係を示す数値「ｋ１１～ｋ１４」を有する。同様に、第２データ２は複数の第２レコードを有し、第２レコードの各々は、第２項目である「項Ｓ，項Ｒ」のそれぞれについての第２項目値（「項Ｓ」についての「Ｓ１，Ｓ２」、「項Ｒ」についての「Ｒ１，Ｒ２」）と、第２項目値間の関係を示す数値「ｋ２１～ｋ２４」を有する。

【００７７】

入力データ距離算出部２２１は、対象項目ごとに、第１データ１内の対象項目に属する対象第１項目値に関する他の第１項目値との関係と、第２データ２内の対象項目に属する対象第２項目値に関する他の第２項目値との関係との類似度を計算する。次に、計算した類似度に基づいて、第１重み情報５，６と第２重み情報７，８とを生成する。第１重み情報５，６は、複数の項目「項Ｓ，項Ｒ」のいずれかに属する複数の変換先項目値「Ｓ’１，Ｓ’２，Ｒ’１，Ｒ’２」のうちの、対象項目に属する対象変換先項目値への対象第１項目値の影響度を示す情報である。第２重み情報７，８は、対象変換先項目値への対象第

２項目値の影響度を示す情報である。

【００７８】

例えば「項Ｓ」が対象項目として選択されたとき、入力データ距離算出部２２１は、第１重み情報５と第２重み情報７とを生成する。このとき第１重み情報５には、第１データ１内の「項Ｓ」に属する項目値「Ｓ１，Ｓ２」それぞれと、第２データ２内の「項Ｓ」に属する項目値「Ｓ１，Ｓ２」それぞれとの対ごとの、他の項目との関係の類似度が維持できるように、重みが設定される。

【００７９】

同様に、第２重み情報７にも、第１データ１内の「項Ｓ」に属する項目値「Ｓ１，Ｓ２」それぞれと、第２データ２内の「項Ｓ」に属する項目値「Ｓ１，Ｓ２」それぞれとの対ごとの、他の項目との関係の類似度が維持できるように、重みが設定される。「項Ｒ」が対象項目として選択されたときに、入力データ距離算出部２２１は、第１重み情報６と第２重み情報８とを生成する。

【００８０】

対象第１項目値と対象第２項目値との類似度を計算する場合、入力データ距離算出部２２１は、例えば複数の項目「項Ｓ，項Ｒ」それぞれについて、初期値が設定された第１重み情報５，６と第２重み情報７，８とを生成する。

【００８１】

次に、入力データ距離算出部２２１は、対象項目以外の項目について生成された他項目第１重み情報と他項目第２重み情報とに基づいて、第１データ１内の対象項目に属する対象第１項目値と、第２データ２内の対象項目に属する対象第２項目値との類似度を計算する。

【００８２】

なお、入力データ距離算出部２２１は、所定の終了条件を満たすまで、複数の項目「項Ｓ，項Ｒ」それぞれを、繰り返し対象項目として特定し、対象項目に対する第１重み情報５，６と第２重み情報７，８とを繰り返し生成してもよい。例えば、入力データ距離算出部２２１は、対象項目以外の項目について生成された他項目第１重み情報と他項目第２重み情報とを用いて、対象項目に属する対象第１項目値それぞれと対象第２項目値それぞれとの類似度を計算し、類似度に応じて対象項目の重み情報を更新する。

【００８３】

具体的には、対象項目が「項Ｓ」であれば、入力データ距離算出部２２１は、「項Ｒ」について生成された第１重み情報６を用いて、第１データ１の項目値「Ｓ１，Ｓ２」それぞれと、第２データ２の項目値「Ｓ１，Ｓ２」それぞれとの類似度を計算する。次に、入力データ距離算出部２２１は、計算した類似度に基づいて、「項Ｓ」についての第１重み情報５と第２重み情報７とを更新する。さらに、入力データ距離算出部２２１は、更新後の第１重み情報５と第２重み情報７を用いて、「項Ｒ」についての第１重み情報６と第２重み情報８とを更新する。

【００８４】

このように第１重み情報５，６と第２重み情報７，８の更新を繰り返すことで、第１類似判断用データ３と第２類似判断用データ４との類似度が向上するように、第１重み情報５，６と第２重み情報７，８が最適化される。

【００８５】

次に、入力データ距離算出部２２１は、複数の項目「項Ｓ，項Ｒ」それぞれについて生成された第１重み情報５，６に基づいて、第１データ１を第１類似判断用データ３に変換する。第１類似判断用データ３は、複数の変換先項目値「Ｓ’１，Ｓ’２，Ｒ’１，Ｒ’２」のうちの異なる項目に属する２以上の変換先項目値間の関係を示す数値「ｋ３１～ｋ３４」が設定された複数の第３レコードを有する。さらに、入力データ距離算出部２２１は、複数の項目「項Ｓ，項Ｒ」それぞれについて生成された第２重み情報７，８に基づいて、第２データ２を第２類似判断用データ４に変換する。第２類似判断用データ４は、複数の変換先項目値「Ｓ’１，Ｓ’２，Ｒ’１，Ｒ’２」のうちの異なる項目に属する２以

上の変換先項目値間の関係を示す数値「ｋ４１～ｋ４４」が設定された複数の第４レコードを有する。

【００８６】

さらに、入力データ距離算出部２２１は、第１類似判断用データ３内の複数の第３レコードに含まれる数値群と、第２類似判断用データ４内の複数の第４レコードに含まれる数値群との類似度を計算する。入力データ距離算出部２２１は、第１重み情報５，６と第２重み情報７，８を繰り返し生成するとき、第１重み情報５，６と第２重み情報７，８を生成するごとに、第１類似判断用データ３の数値群と第２類似判断用データ４の数値群との類似度を計算する。そして、入力データ距離算出部２２１は、計算した類似度の最大値を、第１データ１と第２データ２との類似度と判定する。

【００８７】

このように、第１重み情報５，６と第２重み情報７，８を用いて第１データ１と第２データ２とを変換した上で、類似度を計算することで、精度の高い類似度を算出することができる。すなわち、第１データ１と第２データ２との同一の項目に属する項目値に関する、他の項目値との間の関係の類似度が、その項目に対応する第１重み情報５，６と第２重み情報７，８とに反映されている。これにより、第１データ１と第２データ２との同一の項目に属する項目値に関する、他の項目値との間の関係の類似度が高いほど、変換後の第１類似判断用データ３と第２類似判断用データ４との類似度が高くなる。その結果、類似度の判定精度が向上し、第１データと第２データとの距離についても精度良く算出することができる。

【００８８】

しかも、組み合わせ爆発のような計算量の急激な増加は発生せず、現実的な処理量での類似度計算が可能である。例えば、類似度の計算処理は、行列を用いて以下のように計算できる。

【００８９】

入力データ距離算出部２２１は、特定の項目の項目値と他の項目との関係をベクトルで表現する。そして、入力データ距離算出部２２１は、第１データ１と第２データ２とを、２つの項目値に対応するベクトル間の距離を保持したまま、第１類似判断用データ３と第２類似判断用データ４とに変換する。このとき、入力データ距離算出部２２１は、変換に用いる第１重み情報５，６と第２重み情報７，８とを行列で表す。以下、第１重み情報５，６と第２重み情報７，８とを表す行列を、変換行列と呼ぶ。

【００９０】

入力データ距離算出部２２１は、第１類似判断用データ３と第２類似判断用データ４間の最大類似度を、第１データ１と第２データ２との間の類似度とする。これにより、本質的な関係の構造に基づく類似度を計算でき、第１データ１と第２データ２との間の本質的な関係の構造に基づく距離を計算することができる。

【００９１】

以下に、変換行列を用いた類似度の計算に関する詳細を説明する。上述のように、簡単のため、第１データ１と第２データ２との項目は２つだけとする。入力データ距離算出部２２１は、第１データ１と第２データ２とを、行列X_1，X_2で表す。行列X_1，X_2の各行は、１つ目の項目「項Ｓ」の各項目値「Ｓ１，Ｓ２」に対応し、各列は２つめの項目「項Ｒ」の各変数値「Ｒ１，Ｒ２」に対応する。行列の要素（成分）には、行に対する項目値と列に対応する項目値との関係を示す数値が入る。

【００９２】

なお、入力データ距離算出部２２１は、１つ目の項目「項Ｓ」の項目値の種類の数が第１データ１と第２データ２とで異なる場合には、少ないほうのデータにダミーの項目値を追加して、種類の数を同数にする。入力データ距離算出部２２１は、２つ目の項目「項Ｒ」についても同様に、項目値の種類数を同数に揃える。

【００９３】

入力データ距離算出部２２１は、第１データ１の「項Ｓ」と「項Ｒ」とに関する変換行

列（第１重み情報５，６）を、それぞれ正方行列C_{11}とC_{12}で表す。同様に、入力データ距離算出部２２１は、第２データ２の「項Ｓ」と「項Ｒ」とに関する変換行列（第２重み情報７，８）を、それぞれ正方行列C_{21}とC_{22}で表す。ただし、C_{11}，C_{12}，C_{21}，C_{22}は、いずれも以下の正規直交条件を満たすものとする。

【００９４】

【数１】

$$C_{11}^T C_{11} = C_{21}^T C_{21} = I$$

$$C_{12}^T C_{12} = C_{22}^T C_{22} = I$$

【００９５】

Ｉは対角成分が「１」で残りが「０」の単位行列である。このとき、X_1の列ベクトルをx_{1a}，x_{1b}とする。x_{1a}，x_{1b}は、「項Ｒ」の変数値「ａ」、「ｂ」と「項Ｓ」との関係を表しており、以下の関係を有する。

【００９６】

【数２】

$$\|C_{11}^T x_{1a} - C_{11}^T x_{1b}\|^2 = \|x_{1a} - x_{1b}\|^2$$

【００９７】

すなわち、C_{11}によるX_1の変換は、項目値の他項目との関係を表すベクトル間の距離を変化させない。C_{12}，C_{21}，C_{22}についても同様である。

【００９８】

入力データ距離算出部２２１は、C_{11}とC_{21}の更新では、C_{12}とC_{22}を固定したときの、データ間類似度を最大化するC_{11}とC_{21}として算出する。データ間類似度Ｅ（X_1，X_2）は、以下の式で表される。

【００９９】

【数３】

$$E(X_1, X_2) = \langle C_{11}^T X_1 C_{12}, C_{21}^T X_2 C_{22} \rangle$$

【０１００】

データ間類似度を最大化するC_{11}とC_{21}は、以下に示す特異値分解により算出できる。

【０１０１】

【数４】

$$C_{11} S C_{21}^T = X_1 C_{12} C_{22}^T X_2^T$$

【０１０２】

ただし、Ｓは非負値を持つ正方対角行列である。

【０１０３】

このようにして、行列を用いて効率的にデータ変換を行い、類似度を計算することができる。

【０１０４】

このとき、行列X_1，X_2でそれぞれ表される第１データ１と第２データ２との距離Ｄ（X_1，X_2）は、以下となる。

【０１０５】

【数５】

$$D(X_1, X_2) = \| C_{11}^T X_1 C_{12}, C_{21}^T X_2 C_{22} \|$$

【０１０６】

図１７の例では、第１重み情報５，６および第２重み情報７，８が変換行列で表されている。例えば、第１重み情報５を示す変換行列の第１行・第１列の成分には、第１データ１の「項Ｓ」に属する項目値「Ｓ１」の、「項Ｓ」に属する変換先項目値「Ｓ’１」への

影響を示す重み（ｗ１１）が設定されている。変換行列を用いると、第１データ１の項目ごとの項目値を成分とする行ベクトルに右から変換行列を乗算すれば、変換先項目値を得ることができる。例えば第１データ１の「項Ｓ」に属する項目値を成分とする行ベクトル（Ｓ１，Ｓ２）に、「項Ｓ」に関する第１重み情報５を示す変換行列を右から掛けることで、「項Ｓ」に属する変換先項目値を示す行ベクトル（Ｓ’１，Ｓ’２）が得られる。

【０１０７】

同様に、第１データ１の「項Ｒ」に属する項目値を成分とする行ベクトル（Ｒ１，Ｒ２）に、「項Ｒ」に関する第１重み情報６を示す変換行列を右から掛けることで、「項Ｒ」に属する変換先項目値を示す行ベクトル（Ｒ’１，Ｒ’２）が得られる。第２データ２の「項Ｓ」に属する項目値を成分とする行ベクトル（Ｓ１，Ｓ２）に、「項Ｓ」に関する第２重み情報７を示す変換行列を右から掛けることで、「項Ｓ」に属する変換先項目値を示す行ベクトル（Ｓ’１，Ｓ’２）が得られる。第２データ２の「項Ｒ」に属する項目値を成分とする行ベクトル（Ｒ１，Ｒ２）に、「項Ｒ」に関する第２重み情報８を示す変換行列を右から掛けることで、「項Ｒ」に属する変換先項目値を示す行ベクトル（Ｒ’１，Ｒ’２）が得られる。

【０１０８】

ここで第１データ１と第２データ２とにおける同一レコード内の「項Ｓ」の項目値と「項Ｒ」の項目値の乗算結果が、そのレコードの「数値」の値であるものとする。同様に、第１類似判断用データ３と第２類似判断用データ４における同一レコード内の「項Ｓ」の項目値と「項Ｒ」の項目値の乗算結果が、そのレコードの「数値」の値であるものとする。すると、第１類似判断用データ３と第２類似判断用データ４との「数値」の値を算出できる。例えば第１類似判断用データ３の「Ｓ’１」と「Ｒ’１」との組に対応する数値「ｋ３１」は、以下の通りとなる。

ｋ３１＝Ｓ’１×Ｒ’１

＝（ｗ１１×Ｓ１＋ｗ１２×Ｓ２）×（ｗ２１×Ｒ１＋ｗ２２×Ｒ２）

＝ｗ１１×ｗ２１×Ｓ１×Ｒ１＋ｗ１２×ｗ２１×Ｓ２×Ｒ１＋ｗ１１×ｗ２２×Ｓ１×Ｒ２＋ｗ１２×ｗ２２×Ｓ２×Ｒ２

＝ｗ１１×ｗ２１×ｋ１１＋ｗ１２×ｗ２１×ｋ１２＋ｗ１１×ｗ２２×ｋ１３＋ｗ１２×ｗ２２×ｋ１４

同様にして、第１類似判断用データ３と第２類似判断用データ４との「数値」の他の値（ｋ３１～ｋ３４，ｋ４１～ｋ４４）も算出できる。

【０１０９】

入力データ距離算出部２２１は、第１類似判断用データ３と第２類似判断用データ４の「数値」の各値を比較することで、第１類似判断用データ３と第２類似判断用データ４との類似度を計算する。例えば、入力データ距離算出部２２１は、第１類似判断用データ３の各レコードの数値を成分とするベクトルと、第１類似判断用データ４の各レコードの数値を成分とするベクトルとの内積を計算し、内積の結果を類似度とする。

【０１１０】

このように第１重み情報５，６および第２重み情報７，８を行列で表すことで、類似度を計算することができる。

【０１１１】

なお上記の計算例は、簡単のために第１データ１と第２データ２との項目は２つだけとしているが、項目数がもっと多い場合、行列をテンソルに拡張することで類似度を計算できる。なお、行列は、テンソルの一例である。

【０１１２】

比較対象のデータに対応するテンソルをX_m，X_nとする（ｍ、ｎはデータを識別する整数）。X_m，X_nに含まれる項目数がｋ（ｋは２以上の整数）のとき、変換行列をC_kとすると、データの類似判断用データへの変換は、以下の式で表すことができる。

【０１１３】

【数６】

$$X_n \prod_k \times_k C_k \quad \cdots (5)$$

【０１１４】

式（５）の$\times_k$は、テンソルのモード積を示している。式（５）の結果を用いて、テンソルをX_m，X_n間の距離を、以下の式で表すことができる。

【０１１５】

【数７】

$$E(X_m, X_n) = \|X_m\|_2 + \|X_n\|_2 - 2\langle X_m, X_n \prod_k \times_k C_k \rangle \quad \cdots (6)$$

【０１１６】

式（６）に示す距離を最小にする行列C_kが変換行列となる。ただし、C_kは、以下の正規直交条件を満たすものとする。

【０１１７】

【数８】

$$\begin{cases} C_k^T C_k = I \ (I_{kn} \geq I_{km}) \\ C_k C_k^T = I \ (I_{kn} < I_{km}) \end{cases} \quad \cdots (7)$$

【０１１８】

項目ごとのC_kは、以下の特異値分解を、項目ごとに交互に繰り返し行うことで算出できる。

【０１１９】

【数９】

$$P_k S_k Q_k^T = \left(X_n \prod_{\substack{k'=1,\ldots,K \\ k' \neq k}} \times_{k'} C_{k'} \right)^{(k)T} X_m^{(k)} \quad \cdots (8)$$

【０１２０】

式（８）の(k)は、テンソルを、第k番目の項目を列、その他の項目を行とする行列に変換する操作を表す。式（８）により、行列$P_k S_k Q_k^T$が生成される。そしてP_kとQ_k^Tとを用いて、以下の式により行列C_kが得られる。

【０１２１】

【数１０】

$$C_k = P_k Q_k^T \quad \cdots (9)$$

【０１２２】

このような計算により変換行列を求めデータを変換することで、項目数が３以上であっても現実的な計算量で類似度を計算可能であり、距離を計算することもできる。

【０１２３】

図１８は、変換行列を用いて生成した中間データ間の類似度の計算例を示す図である。入力データ距離算出部２２１は、第１データ１０３１と第１データ１０３２それぞれに対して、「量」以外の変数の数に応じた変換行列１０４１～１０４４を生成する。例えば入力データ距離算出部２２１は、第１データ１０３１に対して、「項Ｓ」に対応する変換行列１０４１と「項Ｒ」に対応する変換行列１０４２とを生成する。同様に入力データ距離算出部２２１は、第２データ１０３２に対して、「項Ｓ」に対応する変換行列１０４３と「項Ｒ」に対応する変換行列１０４４とを生成する。

【０１２４】

変換行列１０４１～１０４４は、正規直交条件を満たす２行２列の行列である。変換行列１０４１の各行には、第１データ１０３１における「項Ｓ」の変数値「Ｓ１」、「Ｓ２」が関連付けられている。変換行列１０４３の各行には、第２データ１０３２における「項Ｓ」の変数値「Ｓ１」、「Ｓ２」が関連付けられている。変換行列１０４１の各列には、中間データ１０５１における「項Ｓ」の変数値「Ｓ’１」、「Ｓ’２」が関連付けられている。変換行列１０４３の各列には、中間データ１０５２における「項Ｓ」の変数値「Ｓ’１」、「Ｓ’２」が関連付けられている。変換行列１０４１，１０４３の各成分には、行方向に関連付けられた変数値「Ｓ１」、「Ｓ２」を、列方向に関連付けられた変数値「Ｓ’１」、「Ｓ’２」へ変換する場合の重みが設定されている。

【０１２５】

変換行列１０４２の各行には、第１データ１０３１における「項Ｒ」の変数値「Ｒ１」、「Ｒ２」が関連付けられている。変換行列１０４４の各行には、第２データ１０３２における「項Ｒ」の変数値「Ｒ１」、「Ｒ２」が関連付けられている。変換行列１０４２の各列には、中間データ１０５１における「項Ｒ」の変数値「Ｒ’１」、「Ｒ’２」が関連付けられている。変換行列１０４４の各列には、中間データ１０５２における「項Ｒ」の変数値「Ｒ’１」、「Ｒ’２」が関連付けられている。変換行列１０４２，１０４４の各成分には、行方向に関連付けられた変数値「Ｒ１」、「Ｒ２」を、列方向に関連付けられた変数値「Ｒ’１」、「Ｒ’２」へ変換する場合の重みが設定されている。

【０１２６】

なお図１８の例では、変換行列１０４１～１０４４に設定されている重みの値を小数点２桁までしか示していないが、実際には小数点２桁よりも下位の桁の値も存在するものとする。

【０１２７】

入力データ距離算出部２２１は、変換行列１０４１，１０４２を用いて、第１データ１０３１を中間データ１０５１に変換する。中間データ１０５１には、「項Ｓ」の変数値と「項Ｒ」の変数値の組み合わせに対応する量が設定されている。中間データ１０５１の「項Ｓ」の変数値には「Ｓ’１」または「Ｓ’２」が設定され、「項Ｒ」の変数値には「Ｒ’１」または「Ｒ’２」が設定されている。

【０１２８】

中間データ１０５１の「量」の値は、「項Ｓ」の変数値と「項Ｒ」の変数値との乗算結果である。入力データ距離算出部２２１は、第１データ１０３１を変換行列１０４１，１０４２で変換することで、中間データ１０５１の「量」の値を算出する。例えば変数値「Ｓ’１」は、重み「－０．６８」×「Ｓ１」＋重み「－０．７３」×「Ｓ２」である。変数値「Ｓ’２」は、重み「－０．７３」×「Ｓ１」＋重み「０．６８」×「Ｓ２」である。変数値「Ｒ’１」は、重み「－０．３２」×「Ｒ１」＋重み「－０．９４」×「Ｒ２」である。変数値「Ｒ’２」は、重み「－０．９４」×「Ｒ１」＋重み「０．３２」×「Ｒ２」である。

【０１２９】

このように、変換行列１０４１，１０４２を用いて、変数値「Ｓ’１」、「Ｓ’２」、「Ｒ’１」、「Ｒ’２」の値を、変数値「Ｓ１」、「Ｓ２」、「Ｒ１」、「Ｒ２」と、それらの重みで表すことができる。すると、変数値「Ｓ’１」または「Ｓ’２」と変数値「Ｒ’１」または「Ｒ’２」とを乗算すると、「Ｓ１×Ｒ１」、「Ｓ２×Ｒ１」、「Ｓ１×Ｒ２」、「Ｓ２×Ｒ２」のいずれを含む項が現れる。例えば「Ｓ１×Ｒ１」は、以下の式で表される。

【０１３０】

Ｓ’１×Ｒ’１＝｛（－０．６８×Ｓ１）＋（－０．７３×Ｓ２）｝×｛（－０．３２×Ｒ１）＋（－０．９４×Ｒ２）｝

＝（－０．６８）×（－０．３２）×Ｓ１×Ｒ１＋（－０．７３）×（－０．３２）×Ｓ２×Ｒ１＋（－０．６８）×（－０．９４）×Ｓ１×Ｒ２＋（－０．７３）×（－０．９４）×Ｓ２×Ｒ２

入力データ距離算出部２２１は、「Ｓ１×Ｒ１」、「Ｓ２×Ｒ１」、「Ｓ１×Ｒ２」、「Ｓ２×Ｒ２」の値として、第１データ１０３１における対応する「量」の値を代入する。図１８の例では、「Ｓ１×Ｒ１＝１」、「Ｓ２×Ｒ１＝０」、「Ｓ１×Ｒ２＝１」、「Ｓ２×Ｒ２＝１」である。その結果、中間データ１０５１における「量」の値が求まる。同様に、入力データ距離算出部２２１は、第２データ１０３２を変換行列１０４３，１０４４で変換して、中間データ１０５２を生成する。

【０１３１】

入力データ距離算出部２２１は、中間データ１０５１，１０５２の間の類似度を計算する。例えば入力データ距離算出部２２１は、中間データ１０５１の「量」の各変数値を成分とするベクトルと、中間データ１０５２の「量」の各変数値を成分とするベクトルとを、長さ「１」に正規化後、内積を計算する。そして入力データ距離算出部２２１は、内積の結果を、中間データ１０５１，１０５２間の類似度とする。

【０１３２】

このようにして計算される中間データ１０５１，１０５２間の類似度は、変換行列１０４１～１０４４に設定されている重みに依存する。そこで入力データ距離算出部２２１は、類似度が高くなるように変換行列１０４１～１０４４を更新する。変換行列１０４１～１０４４の更新は、「項Ｓ」の変換行列１０４１，１０４３の更新と、「項Ｒ」の変換行列１０４２，１０４４の更新とが交互に行われる。

【０１３３】

図１９は、「項Ｓ」の変換行列の更新例を示す図である。「項Ｓ」の変換行列１０４１，１０４３を更新する場合、入力データ距離算出部２２１は、「項Ｓ」の変数を固定とし、「項Ｓ」以外の変数を変換して中間データ１０５３，１０５４を生成する。図１９の例では、入力データ距離算出部２２１は、「項Ｒ」の変数値「Ｒ１」「Ｒ２」を変換行列１０４２、１０４４を用いて変換し、中間データ１０５３、１０５４を生成している。中間データ１０５３，１０５４の「量」の値は、「Ｓ１」または「Ｓ２」と「Ｒ’１」または「Ｒ’２」との乗算結果である。例えば第１データ１０３１の中間データ１０５３における「Ｓ１×Ｒ’１」は、変換行列１０４２に示される重みを用いて、「（－０．３２）×Ｓ１×Ｒ１＋（－０．９４）×Ｓ１×Ｒ２」と表される。第１データ１０３１に基づいて、「Ｓ１×Ｒ１」と「Ｓ１×Ｒ２」とに値を設定すれば、「Ｓ１×Ｒ’１」の値が得られる。

【０１３４】

第１データ１０３１と第２データ１０３２の中間データ１０５３，１０５４が生成されると、入力データ距離算出部２２１は、中間データ１０５３における「Ｓ１」、「Ｓ２」それぞれと、中間データ１０５４における「Ｓ１」、「Ｓ２」それぞれとの類似度を計算し、類似度行列１０６１を生成する。類似度行列１０６１の各行には、第１データ１０３１の「項Ｓ」の変数値が関連付けられており、類似度行列１０６１の各列には、第１データ１０３２の「項Ｓ」の変数値が関連付けられている。類似度行列１０６１の成分には、その成分が設定された行の変数値と列の変数値との類似度が設定されている。

【０１３５】

例えば入力データ距離算出部２２１は、「項Ｓ」の各変数値それぞれについて、他の「項Ｒ」の各変数値との関係を示すベクトルを生成する。具体的には、入力データ距離算出部２２１は、中間データ１０５３の「Ｓ１」について、「Ｒ’１」と「Ｒ’２」とのそれぞれとの関係を示す「量」の値を成分とするベクトル$v1_1$を生成する。同様に入力データ距離算出部２２１は、中間データ１０５３の「Ｓ２」について、ベクトル$v2_1$を生成する。入力データ距離算出部２２１は、中間データ１０５４の「Ｓ１」について、ベクトル$v1_2$を生成する。入力データ距離算出部２２１は、中間データ１０５４の「Ｓ２」について、ベクトル$v2_2$を生成する。

【０１３６】

入力データ距離算出部２２１は、ベクトル$v1_1$とベクトル$v1_2$との内積を、第１デー

タ１０３１の「Ｓ１」と第２データ１０３２の「Ｓ１」との類似度として、類似度行列１０６１に設定する。入力データ距離算出部２２１は、ベクトルｖ1_1とベクトルｖ2_2との内積を、第１データ１０３１の「Ｓ１」と第２データ１０３２の「Ｓ２」との類似度として、類似度行列１０６１に設定する。入力データ距離算出部２２１は、ベクトルｖ2_1とベクトルｖ1_2との内積を、第１データ１０３１の「Ｓ２」と第２データ１０３２の「Ｓ１」との類似度として、類似度行列１０６１に設定する。入力データ距離算出部２２１は、ベクトルｖ2_1とベクトルｖ2_2との内積を、第１データ１０３１の「Ｓ２」と第２データ１０３２の「Ｓ２」との類似度として、類似度行列１０６１に設定する。

【０１３７】

入力データ距離算出部２２１は、このようにして生成した類似度行列１０６１に基づいて、第１データ１０３１の「項Ｓ」変換用の変換行列１０４１ａと第１データ１０３２の「項Ｓ」変換用の変換行列１０４３ａとを生成する。例えば入力データ距離算出部２２１は、変換行列１０４１ａ，１０４３ａからＳ’１、Ｓ’２を消去して１つの行列にしたときに類似度行列１０６１に最も類似するような、変換行列１０４１ａ，１０４３ａを生成する。具体的には、入力データ距離算出部２２１は、類似度行列１０６１を特異値分解し、変換行列１０４１ａ，１０４３ａを生成する。

【０１３８】

類似度行列１０６１は、第１データ１０３１の「項Ｓ」変換用の変換行列１０４１を、生成した変換行列１０４１ａに更新する。また類似度行列１０６１は、第２データ１０３２の「項Ｓ」変換用の変換行列１０４３を、生成した変換行列１０４３ａに更新する。

【０１３９】

このようにして、「項Ｓ」を固定して他の変数を変換することで、「項Ｓ」の変換行列が更新される。次に入力データ距離算出部２２１は、「項Ｒ」を固定して他の変数を変換することで、「項Ｒ」の変換行列を更新する。入力データ距離算出部２２１は、各変数の変換行列を更新したら、更新後の変換行列を用いて、第１データ１０３１と第１データ１０３２との中間データを生成し、中間データ間の類似度を計算する。入力データ距離算出部２２１は、例えば中間データ間の類似度が収束するまで、変換行列の更新を繰り返し行う。これにより中間データ間の類似度の最大値を得る変換行列が生成される。そして入力データ距離算出部２２１は、中間データ間の類似度の最大値を、第１データ１０３１と第２データ１０３２との類似度とする。

【０１４０】

図１８、図１９に示した処理の手順をフローチャートで表すと図２０のようになる。

【０１４１】

図２０は、類似度計算処理の手順の一例を示すフローチャートである。以下、図２０に示す処理をステップ番号に沿って説明する。類似度計算処理は、類似度の比較対象となる２つのデータが入力されたときに実行される。

【０１４２】

［ステップＳ１０１］入力データ距離算出部２２１は、変換行列を初期化する。例えば入力データ距離算出部２２１は、分類対象のデータの「量」以外の変数ごとに、変換行列を生成する。生成される変換行列は、対応する変数に含まれる変数値の数（同一の値の変数値は１つと数える）分の行と列とを有する正方行列である。変換行列の成分には、正規直交条件を満たしていれば、ランダムな値を設定することができる。例えば入力データ距離算出部２２１は、変換行列内のいくつかの成分の値をランダムに決定し、正規直交条件を満たすように他の成分の値を決定する。

【０１４３】

［ステップＳ１０２］入力データ距離算出部２２１は、生成した変換行列を用いて、比較対象のデータそれぞれから中間データを生成する。

【０１４４】

［ステップＳ１０３］入力データ距離算出部２２１は、中間データ間の類似度を算出する。入力データ距離算出部２２１は、算出した類似度をメモリに一時的に保存する。

【０１４５】
［ステップＳ１０４］入力データ距離算出部２２１は、比較対象のデータの変数を１つ選択する。
【０１４６】
［ステップＳ１０５］入力データ距離算出部２２１は、比較対象のデータそれぞれの変数値間の類似度を示す類似度行列を生成する。例えば入力データ距離算出部２２１は、比較対象のデータそれぞれについて、選択した変数以外の変数を変換行列で変換した中間データを生成する。そして入力データ距離算出部２２１は、中間データに示される変数値の量の値と、他の中間データに示される変数値の量の値との類似度を、それらの２つの変数値間の類似度を示す成分として、類似度行列に設定する。
【０１４７】
［ステップＳ１０６］入力データ距離算出部２２１は、類似度行列に基づいて、選択した変数についての新たな変換行列を生成する。
【０１４８】
［ステップＳ１０７］入力データ距離算出部２２１は、すべての変数を選択したか否かを判断する。すべての変数の選択が選択済みとなった場合、処理がステップＳ１０８に進められる。未選択の変数があれば、処理がステップＳ１０４に進められる。
【０１４９】
［ステップＳ１０８］入力データ距離算出部２２１は、各変数について新たに生成した変換行列を用いて、比較対象のデータごとの中間データを生成する。
【０１５０】
［ステップＳ１０９］入力データ距離算出部２２１は、ステップＳ１０９で生成した中間データ間の類似度を算出する。
【０１５１】
［ステップＳ１１０］入力データ距離算出部２２１は、処理の終了条件が満たされたか否かを判断する。処理の終了条件とは、例えば類似度が収束したか、またはステップＳ１０４～Ｓ１１０のループを所定回数以上繰り返したことである。処理の終了条件が満たされた場合、類似度計算処理が終了する。処理の終了条件が満たされていない場合、入力データ距離算出部２２１は、変数の選択状態を未選択に初期化して、処理をステップＳ１０４に進める。
【０１５２】
このような手順で比較対象のデータ間の類似度を計算することができる。以下、図２１～２４を参照して、類似度計算の具体例について説明する。
【０１５３】
図２１は、初期状態の変換行列を用いた中間データの生成例を示す図である。図２１の例では、第１データ１０３１と第２データ１０３２とが比較対象のデータである。まず、第１データ１０３１の「項Ｓ」の変換行列１０４１と「項Ｒ」の変換行列１０４２とが初期化され、初期状態の変換行列１０４１，１０４２を用いて、第１データ１０３１が中間データ１０５１に変換される。同様に、第２データ１０３２の「項Ｓ」の変換行列１０４３と「項Ｒ」の変換行列１０４４とが初期化され、初期状態の変換行列１０４３，１０４４を用いて、第２データ１０３２が中間データ１０５２に変換される。そして、第１データ１０３１の中間データ１０５１と第２データ１０３２の中間データ１０５２との類似度が算出される。図１１の例では、類似度が「０．４０」である。
【０１５４】
次に、変数「項Ｓ」が選択されたものとする。「項Ｓ」が選択されると、「項Ｓ」の変換行列が更新される。
【０１５５】
図２２は、「項Ｓ」の変換行列の更新例を示す図である。第１データ１０３１について、「項Ｒ」用の変換行列１０４２を用いて、「項Ｓ」以外の変数値を変換した中間データ１０５３が生成される。同様に第２データ１０３２について、「項Ｒ」用の変換行列１０

４４を用いて、「項Ｓ」以外の変数値を変換した中間データ１０５４が生成される。次に、生成された２つの中間データ１０５３，１０５４それぞれの変数値間の類似度を示す類似度行列１０６１が生成される。そして類似度行列１０６１に基づいて、掛け合わせることで類似度行列１０６１と近似した値を得ることができる２つの変換行列１０４１ａ，１０４３ａが生成される。

【０１５６】

次に、変数「項Ｒ」が選択されたものとする。「項Ｒ」が選択されると、「項Ｒ」の変換行列が更新される。

【０１５７】

図２３は、「項Ｒ」の変換行列の更新例を示す図である。第１データ１０３１について、「項Ｓ」用の変換行列１０４１ａを用いて、「項Ｒ」以外の変数値を変換した中間データ１０５５が生成される。同様に第２データ１０３２について、「項Ｓ」用の変換行列１０４３ａを用いて、「項Ｒ」以外の変数値を変換した中間データ１０５６が生成される。次に、生成された２つの中間データ１０５５，１０５６それぞれの変数値間の類似度を示す類似度行列１０６２が生成される。そして類似度行列１０６２に基づいて、掛け合わせることで類似度行列１０６２と近似した値を得ることができる２つの変換行列１０４２ａ，１０４４ａが生成される。

【０１５８】

「項Ｓ」と「項Ｒ」とのそれぞれについて、変換行列の更新が終了すると、更新後の変換行列１０４１ａ，１０４２ａ，１０４３ａ，１０４４ａに基づいて、中間データが生成される。

【０１５９】

図２４は、更新後の変換行列を用いた中間データの生成例を示す図である。更新後の変換行列１０４１ａ，１０４２ａを用いて、第１データ１０３１から中間データ１０５３が生成される。同様に、更新後の変換行列１０４３ａ，１０４４ａを用いて、第２データ１０３２から中間データ１０５４が生成される。そして、生成された中間データ１０５３，１０５４間の類似度が算出される。図２４の例では、類似度は「０．９１」である。

【０１６０】

更新後の変換行列１０４１ａ，１０４２ａ，１０４３ａ，１０４４ａを用いて生成した中間データ１０５３，１０５４間の類似度は、更新前の変換行列１０４１～１０４４を用いて生成した中間データ１０５１，１０５２間の類似度（図１８参照）よりも高くなっている。すなわち変換行列１０４１～１０４４を更新したことにより、生成される中間データ間の類似度が高まっている。更新後の変換行列１０４１ａ，１０４２ａ，１０４３ａ，１０４４ａに対して、さらに更新処理を行えば、中間データ間の類似度をさらに高めることができる。ただし更新処理を何度も繰り返すと、中間データ間の類似度の上昇度合いが鈍化し、ある程度の類似度に収束する。

【０１６１】

入力データ距離算出部２２１は、例えば、類似度の上昇が所定値以下になったとき、最後に算出した中間データ間の類似度を、比較対象の第１データ１０３１と第２データ１０３２との類似度に決定する。入力データ距離算出部２２１は、決定した類似度から距離を算出し、入力データ距離記憶部２２２に格納する。

【０１６２】

このように、変換行列を用いて生成した中間データ間の類似度により、複数のデータ間の類似度を決定するようにしたことで、類似度の判定精度が向上し、データ間の距離についての精度も向上する。

【０１６３】

以下に、データに含まれる変数値の並べ替えのみによる類似度の計算例である比較例の図２５および図２６と、変換行列を用いて生成した中間データ間の類似度の計算例である図２７および図２８とについて説明する。

【０１６４】

図２５は、類似度計算の比較例を示す第１の図である。図２５に示す第１データ１０７１と第２データ１０７２との類似度を計算する場合を想定する。人やモノの間の関係のしかたを分類するとき、第１データ１０７１を採取した期間に「Ｓ１」の装置が担っていた役割を、第２データ１０７２を採取した期間では「Ｓ２」の装置が担っている可能性がある。そこで図２５の例では、通信元ホストや通信先ホストを、別の変数値に対応付けて、各データ内のレコードの並べ替えを行っている。

【０１６５】

並べ替えにより、変換データ１０７３，１０７４が生成される。２つの変換データ１０７３，１０７４は、通信元ホスト、通信先ホスト、ポートの関係を示す変数値の組み合わせの順番が統一されている。図２５の例では、変換データ１０７３，１０７４の最上位には、「Ｓ’１」、「Ｒ’１」、「Ｐ’１」の組み合わせを示すレコードが登録され、その次に「Ｓ’１」、「Ｒ’１」、「Ｐ’２」の組み合わせを示すレコードが登録されている。

【０１６６】

このように変換データ１０７３，１０７４内に所定の順番で並べられた各レコードの量の値を比較することで、変換データ１０７３，１０７４間の類似度を算出できる。例えば、量の値を成分とするレベクトル間の内積が、類似度とされる。この場合、第１データ１０７１と第２データ１０７２との各変数値に、変換データ１０７３，１０７４のどの変数値を対応付けるかにより、類似度が変わってくる。そのため、対応付けのすべてのパタンについて変換データ１０７３，１０７４を生成し、類似度の最大化が図られる。そして、変換データ１０７３，１０７４から得られる類似度の最大値が、第１データ１０７１と第２データ１０７２との類似度と判定される。

【０１６７】

図２６は、類似度計算の比較例を示す第２の図である。図２５に示した方法で類似度を計算したときの第１データ１０７１と第２データ１０７２との類似度が「０．８９」である。同じ方法で、第１データ１０７１と第３データ１０７５との類似度を計算すると、同じく「０．８９」となる。

【０１６８】

ここで、第１データ１０７１と第２データ１０７２における通信先ホストとポートとの関係をみると、「量」の値が「１」以上のレコードに、｛Ｒ１，Ｐ１｝または｛Ｒ２，Ｐ２｝の組み合わせしか含まれていないことが分かる。それに対して第３データ１０７５では、「量」の値が「１」以上のレコードのなかに、｛Ｒ１，Ｐ２｝の組み合わせを含むものがある。そうすると、第１データ１０７１と第２データ１０７２とは、通信元ホストが分離・併合された前後での通信ログというだけで、類似の事象に関する通信ログである可能性が高い。それに対して、第３データ１０７５は、別の事象に関する通信ログであると考えられる。

【０１６９】

しかし、図２５に示した方法で類似度を計算すると、図２６に示すように、第１データ１０７１から見たとき、第２データ１０７２と第３データ１０７５とのいずれとの間も類似度が同じとなる。すなわち、正しく類似度が計算されていない。

【０１７０】

次に第１データ１０７１と第２データ１０７２との類似度、および第１データ１０７１と第３データ１０７５との類似度を、第２の実施形態に係る方法で計算した場合について、図２７，図２８を参照して説明する。

【０１７１】

図２７は、入力データ距離算出部２２１による類似度計算例を示す第１の図である。図２７には、第１データ１０７１と第２データ１０７２との類似度の計算例を示している。第１データ１０７１について、通信元ホスト、通信先ホスト、ポートそれぞれに対応する変換行列１０８１－２～１０８３－２が生成されている。これらの変換行列１０８１－２～１０８３－２を用いて、第１データ１０７１が中間データ１０９１－２に変換されてい

る。また第２データ１０７２について、通信元ホスト、通信先ホスト、ポートそれぞれに対応する変換行列１０８４～１０８６が生成されている。これらの変換行列１０８４～１０８６を用いて、第２データ１０７２が中間データ１０９２に変換されている。第１データ１０７１の中間データ１０９１と第２データ１０７２の中間データ１０９２との類似度は、「０．９７」である。

【０１７２】

図２８は、入力データ距離算出部２２１による類似度計算例を示す第２の図である。図２８には、第１データ１０７１と第３データ１０７５との類似度の計算例を示している。第１データ１０７１について、通信元ホスト、通信先ホスト、ポートそれぞれに対応する変換行列１０８１－３～１０８３－３が生成されている。これらの変換行列１０８１－３～１０８３－３を用いて、第１データ１０７１が中間データ１０９１－３に変換されている。第３データ１０７５について、通信元ホスト、通信先ホスト、ポートそれぞれに対応する変換行列１０８７～１０８９が生成されている。これらの変換行列１０８７～１０８９を用いて、第３データ１０７５が中間データ１０９３に変換されている。第１データ１０７１の中間データ１０９１－３と第３データ１０７５の中間データ１０９３との類似度は、「０．９４」である。

【０１７３】

図２７と図２８の類似度の計算結果から、第１データ１０７１は、第３データ１０７５よりも第２データ１０７２に類似していることが分かる。すなわち、変数値間の関係を正しく反映させた類似度計算により、類似度の計算制度が向上している。

【０１７４】

しかも図２５、図２６に示したような方法で類似度の精度を上げようとすると、対応付けのすべてのパタンについて類似度を計算することとなり、計算量が膨大となる。

【０１７５】

例えば、３項目の人またはものがあり、各項目の種類数がそれぞれ「Ａ，Ｂ，Ｃ」（Ａ，Ｂ，Ｃは１以上の整数）であるものとする。このとき、図２５、図２６に示した方法で類似度を計算すると、「Ａ！Ｂ！Ｃ！」の数の組み合わせパタンについて類似度計算を行うこととなる。それに対して、入力データ距離算出部２２１による手法では、（$A^2+B^2+C^2$）ＡＢＣに比例する計算量となる。これは「Ａ，Ｂ，Ｃ」がそれぞれ「１０，１０，１０」なら、約１６０，０００，０００，０００，０００倍高速となることを意味する。

【０１７６】

以上のように、入力データ距離算出部２２１による類似度計算方法によれば、類似する事象がログ生成の過程で異なる状態で記録された場合でも、重みづけによる変換により、精度の高い類似度の判定を、効率的に実行することができ、精度の高い距離の判定を行うことができる。

【０１７７】

〔その他の実施の形態〕

第２の実施形態では、単位期間ごとの通信ログ間の類似度および距離を計算する例を示したが、同じ技術により、他の様々な情報の類似度および距離を計算可能である。

【０１７８】

以上、実施の形態を例示したが、実施の形態で示した各部の構成は同様の機能を有する他のものに置換することができる。また、他の任意の構成物や工程が付加されてもよい。さらに、前述した実施の形態のうちの任意の２以上の構成（特徴）を組み合わせたものであってもよい。

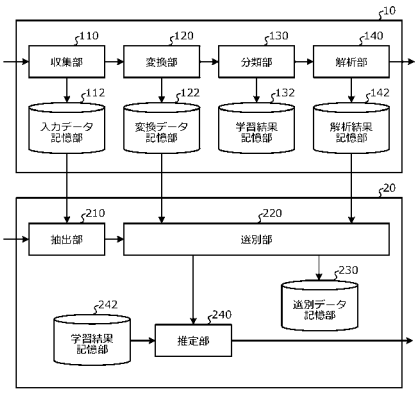
【符号の説明】

【０１７９】

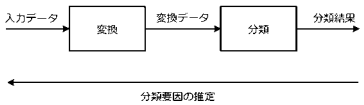
１０　分類装置
１１０　収集部
１１２　入力データ記憶部

１２０　変換部
１２２　変換データ記憶部
１３０　分類部
１３２　学習結果記憶部
１４０　解析部
１４２　解析結果記憶部
２０　要因推定装置
２１　入力データの特徴空間
２２　変換データの特徴空間
２４　近似識別線
２５－１～２５－４　説明ベクトル
２６　点ｕにおける識別要因
２１０　抽出部
２２０　選別部
２２１　入力データ距離算出部
２２２　入力データ距離記憶部
２２３　変換データ距離算出部
２２４　変換データ距離記憶部
２２５　対象判定部
２３０　選別データ記憶部
２４０　推定部
２４２　学習結果記憶部

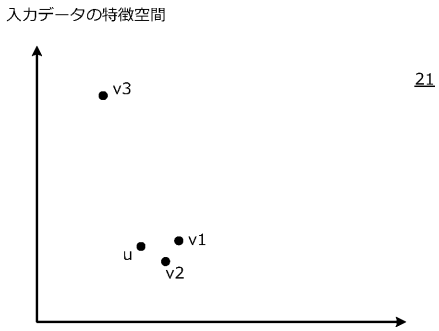
【図１】

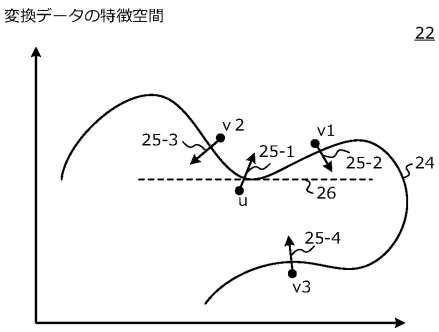


【図２】

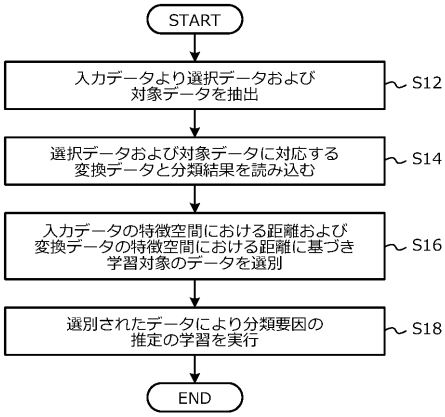


【図３】

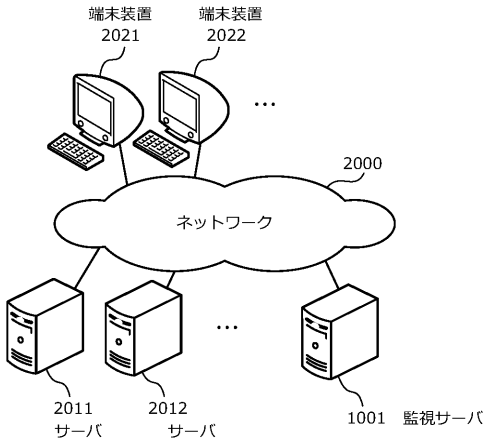




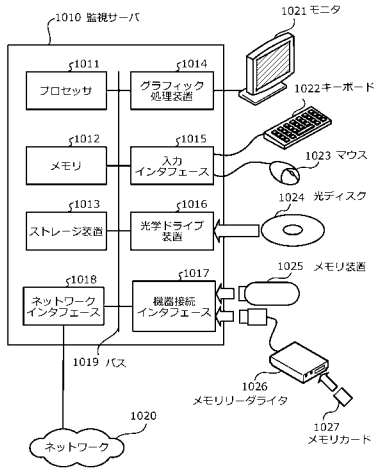
【図４】



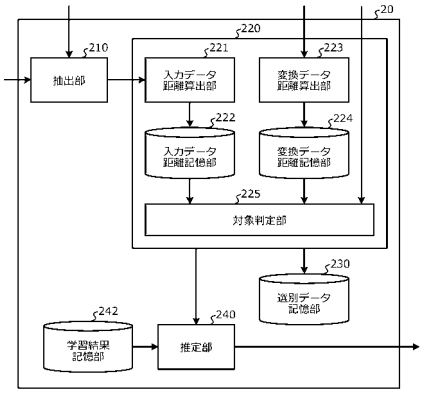
【図５】



【図６】



【図７】



【図８Ａ】

803 順序付け

801 入力データ

通信元ホスト	通信先ホスト	ポート	時
S3	R1	P1	1
S3	R2	P1	1
S3	R3	P1	1
S1	R3	P3	2
S2	R1	P1	1

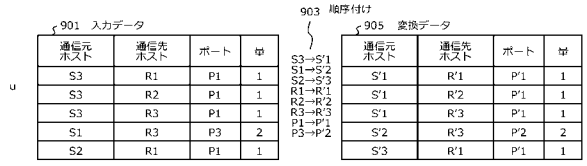
S3→S'1
S1→S'2
S2→S'3
R1→R'1
R2→R'2
R3→R'3
P1→P'1
P3→P'2

805 中間データ

通信元ホスト	通信先ホスト	ポート	時
S'1	R'1	P'1	1
S'1	R'2	P'1	1
S'1	R'3	P'1	1
S'2	R'3	P'2	2
S'3	R'1	P'1	1

【図８Ｂ】

【図９Ａ】

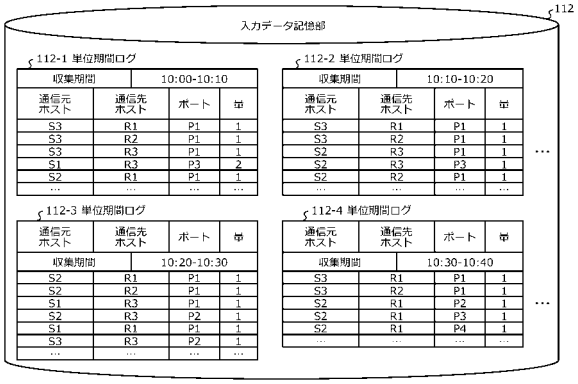


【図９Ｂ】

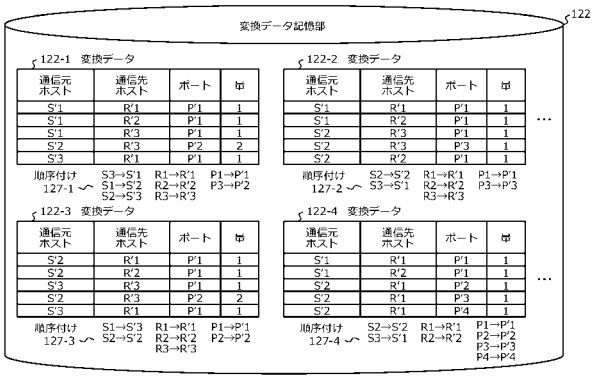
【図１０】

	v1	v2	v3
入力データ特徴空間	5	4	9
変換データ特徴空間	9	8	9
合計	14	12	18

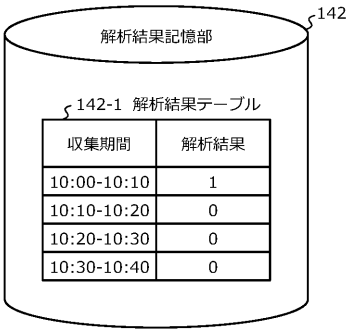
【図１１】



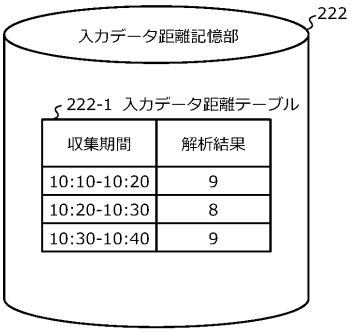
【図１２】



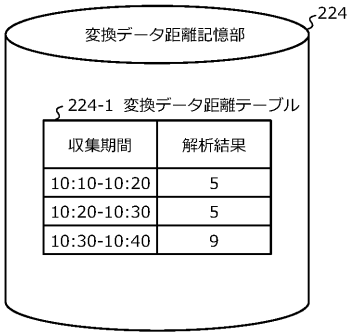
【図１３】



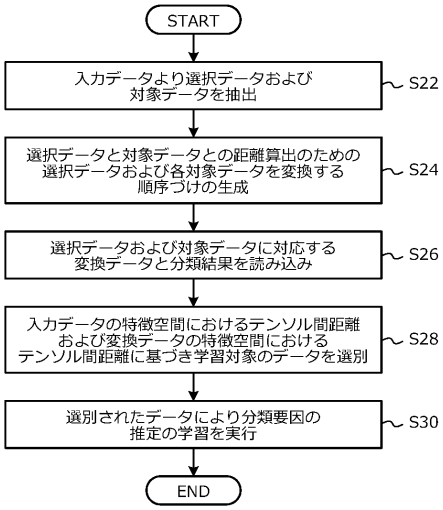
【図１４】



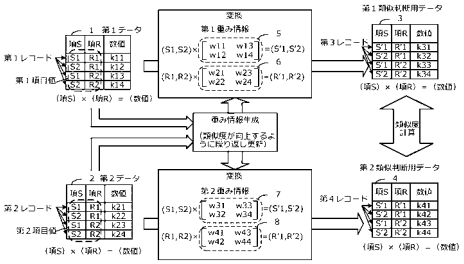
【図１５】



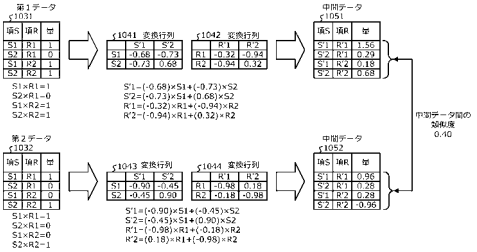
【図１６】



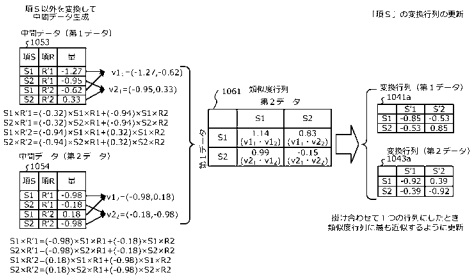
【図１７】



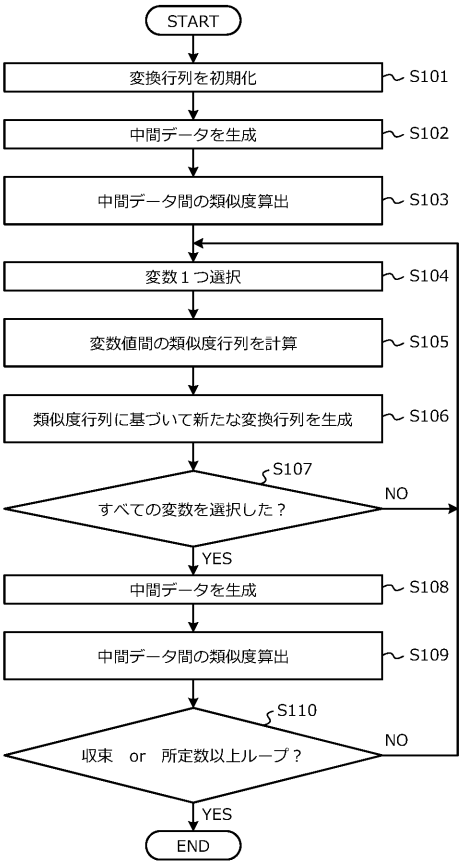
【図１８】



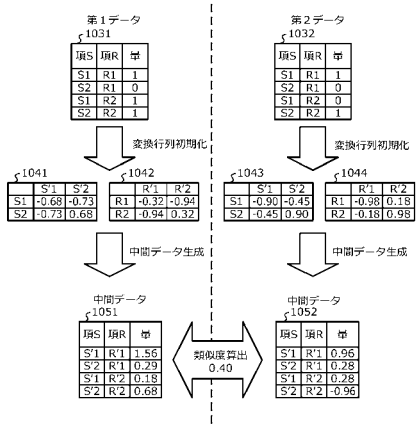
【図１９】



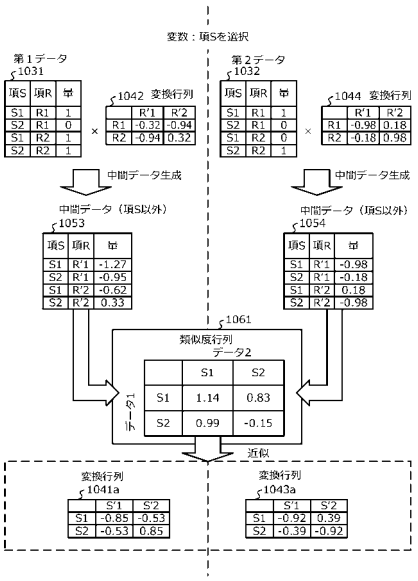
【図２０】



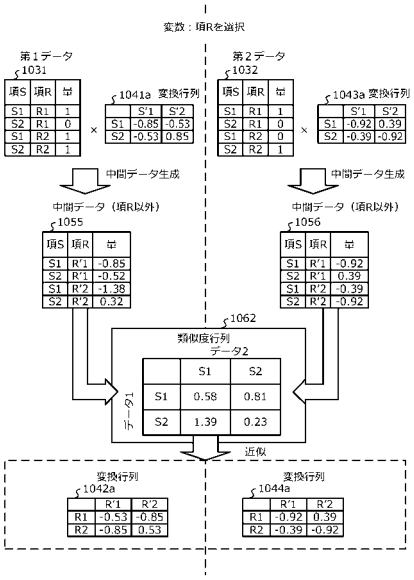
【図２１】



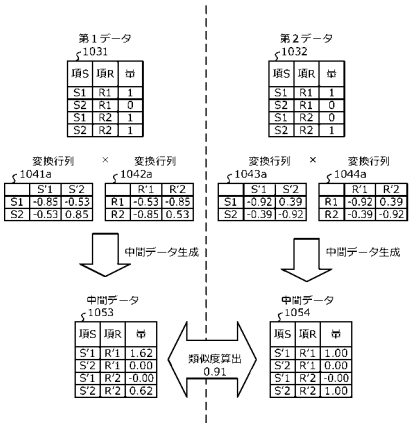
【図２２】



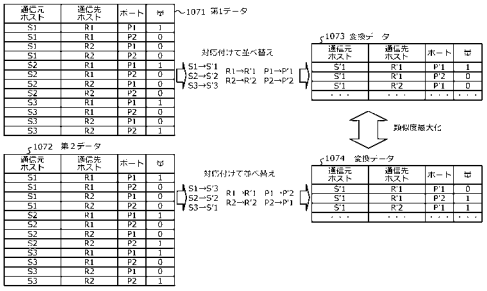
【図２３】



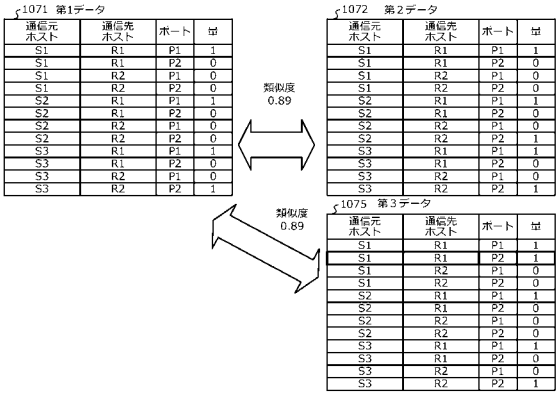
【図２４】



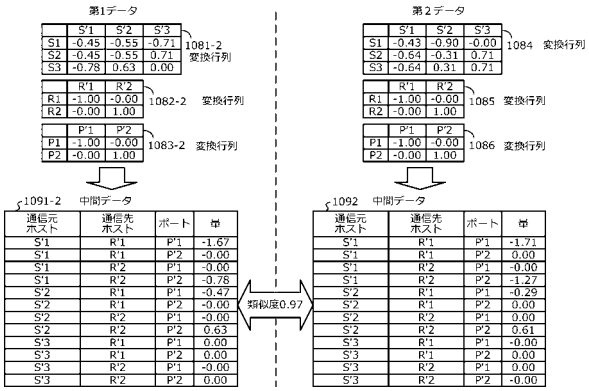
【図２５】



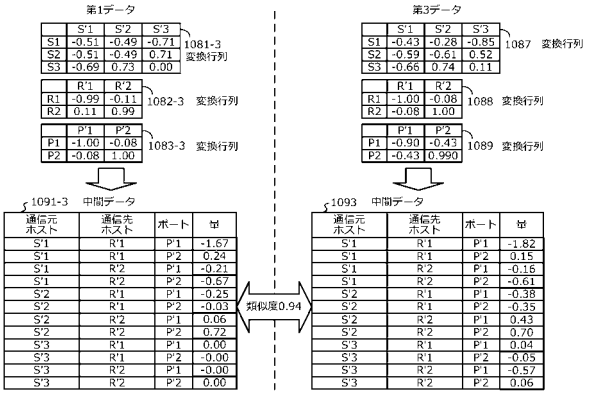
【図２６】



【図２７】



【図２８】


第1データ
S'1 S'2 S'3
S1 -0.51 -0.49 -0.71
S2 -0.51 -0.49 0.71
S3 -0.69 0.73 0.00
1081-3 変換行列
R'1 R'2
R1 -0.99 -0.11
R2 0.11 0.99
1082-3 変換行列
P'1 P'2
P1 -1.00 -0.08
P2 -0.08 1.00
1083-3 変換行列
1091-3 中間データ
通信元ホスト 通信先ホスト ポート 値
S'1 R'1 P'1 -1.67
S'1 R'1 P'2 0.24
S'1 R'2 P'1 -0.21
S'1 R'2 P'2 -0.67
S'2 R'1 P'1 -0.25
S'2 R'1 P'2 -0.03
S'2 R'2 P'1 0.06
S'2 R'2 P'2 0.72
S'3 R'1 P'1 0.00
S'3 R'1 P'2 -0.00
S'3 R'2 P'1 -0.00
S'3 R'2 P'2 0.00
類似度0.94
第3データ
S'1 S'2 S'3
S1 -0.43 -0.28 -0.85
S2 -0.59 -0.61 0.52
S3 -0.66 0.74 0.11
1087 変換行列
R'1 R'2
R1 -1.00 -0.08
R2 -0.08 1.00
1088 変換行列
P'1 P'2
P1 -0.90 -0.43
P2 -0.43 0.990
1089 変換行列
1093 中間データ
通信元ホスト 通信先ホスト ポート 値
S'1 R'1 P'1 -1.82
S'1 R'1 P'2 0.15
S'1 R'2 P'1 -0.16
S'1 R'2 P'2 -0.61
S'2 R'1 P'1 -0.38
S'2 R'1 P'2 -0.35
S'2 R'2 P'1 0.43
S'2 R'2 P'2 0.70
S'3 R'1 P'1 0.04
S'3 R'1 P'2 -0.05
S'3 R'2 P'1 -0.57
S'3 R'2 P'2 0.06