

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7073286号
(P7073286)

(45)発行日 令和4年5月23日(2022.5.23)

(24)登録日 令和4年5月13日(2022.5.13)

(51)国際特許分類	F I			
G 0 6 N	20/00 (2019.01)	G 0 6 N	20/00	
G 0 6 N	3/08 (2006.01)	G 0 6 N	3/08	1 4 0

請求項の数 13 (全18頁)

(21)出願番号	特願2019-2436(P2019-2436)	(73)特許権者	000005108
(22)出願日	平成31年1月10日(2019.1.10)		株式会社日立製作所
(65)公開番号	特開2020-112967(P2020-112967)		東京都千代田区丸の内一丁目6番6号
	A)	(74)代理人	110001678
(43)公開日	令和2年7月27日(2020.7.27)		特許業務法人藤央特許事務所
審査請求日	令和3年4月21日(2021.4.21)	(72)発明者	但馬 慶行 東京都千代田区丸の内一丁目6番6号 株式会社日立製作所内
		(72)発明者	河野 洋平 東京都千代田区丸の内一丁目6番6号 株式会社日立製作所内
		審査官	金木 陽一

最終頁に続く

(54)【発明の名称】 データ生成装置、予測器学習装置、データ生成方法、及び学習方法

(57)【特許請求の範囲】

【請求項1】

データ集合を生成するデータ生成装置であって、
訓練データ集合の各元の入力及び前記訓練データ集合に関する情報の少なくとも一方に基づいて、前記元を変形するための摂動集合を生成する摂動生成部と、
前記訓練データ集合及び前記摂動集合から、前記訓練データ集合と異なる新たな疑似データ集合を生成する疑似データ合成部と、
前記訓練データ集合と前記疑似データ集合との分布間距離又はそれに関する推定量と、前記摂動集合から得られる訓練データに対する疑似データの摂動の大きさとを算出する評価部と、
前記訓練データ集合と前記疑似データ集合との分布間距離を近づけ、摂動の大きさ又は期待値が予め定めた目標値となるように、前記摂動生成部が前記摂動集合の生成に使用するパラメータを更新するパラメータ更新部とを備えることを特徴とするデータ生成装置。

【請求項2】

請求項1に記載のデータ生成装置であって、
前記摂動生成部は、前記訓練データ集合の各元の入力又は前記訓練データ集合に関する情報に加えて、前記訓練データ集合の各元の出力又はそれに関する情報に基づいて前記摂動集合を生成することを特徴とするデータ生成装置。

【請求項3】

請求項1に記載のデータ生成装置であって、

前記摂動生成部は、前記訓練データ集合の各元の入力又は前記訓練データ集合に関する情報に加えて、前記訓練データ集合の入力に関する確率密度関数の推定量に基づいて前記摂動集合を生成することを特徴とするデータ生成装置。

【請求項 4】

請求項 1 に記載のデータ生成装置であって、

前記摂動生成部は、前記摂動集合の事後分布を表すパラメトリックな分布の母数を生成することによって、前記摂動集合を生成することを特徴とするデータ生成装置。

【請求項 5】

請求項 1 に記載のデータ生成装置であって、

前記摂動生成部が使用するパラメータ値又はその範囲を入力可能なインターフェース画面の表示データを生成することを特徴とするデータ生成装置。

10

【請求項 6】

請求項 1 に記載のデータ生成装置であって、

前記訓練データ集合の各元と前記疑似データ集合の各元とが表された散布図の表示データを生成することを特徴とするデータ生成装置。

【請求項 7】

計算機がデータ集合を生成するデータ生成方法であって、

前記計算機は、所定の演算処理を実行する演算装置と、前記演算装置がアクセス可能な記憶装置とを有し、

前記データ生成方法は、

前記演算装置が、訓練データ集合の各元の入力及び前記訓練データ集合に関する情報の少なくとも一方に基づいて、前記元を変形するための摂動集合を生成する摂動生成手順と、
前記演算装置が、前記訓練データ集合及び前記摂動集合から、前記訓練データ集合と異なる新たな疑似データ集合を生成する疑似データ合成手順と、

前記演算装置が、前記訓練データ集合と前記疑似データ集合との分布間距離又はそれに関する推定量と、前記摂動集合から得られる訓練データに対する疑似データの摂動の大きさとを算出する評価手順と、

前記訓練データ集合と前記疑似データ集合との分布間距離を近づけ、摂動の大きさ又は期待値が予め定めた目標値となるように、前記摂動生成手順において前記摂動集合の生成に使用するパラメータを更新するパラメータ更新手順とを含むことを特徴とするデータ生成方法。

20

【請求項 8】

請求項 7 に記載のデータ生成方法であって、

前記摂動生成手順では、前記演算装置が、前記訓練データ集合の各元の入力又は前記訓練データ集合に関する情報に加えて、前記訓練データ集合の各元の出力又はそれに関する情報に基づいて前記摂動集合を生成することを特徴とするデータ生成方法。

30

【請求項 9】

請求項 7 に記載のデータ生成方法であって、

前記摂動生成手順では、前記演算装置が、前記摂動集合の事後分布を表すパラメトリックな分布の母数を生成することによって、前記摂動集合を生成することを特徴とするデータ生成方法。

40

【請求項 10】

請求項 7 に記載のデータ生成方法であって、

前記演算装置が、前記摂動生成手順で使用されるパラメータ値又はその範囲を入力可能なインターフェース画面の表示データを生成する手順を含むことを特徴とするデータ生成方法。

【請求項 11】

請求項 7 に記載のデータ生成方法であって、

前記演算装置が、前記訓練データ集合の各元と前記疑似データ集合の各元とが表された散布図の表示データを生成する手順を含むことを特徴とするデータ生成方法。

50

【請求項 1 2】

計算機がデータ集合を学習する学習方法であって、

前記計算機は、所定の演算処理を実行する演算装置と、前記演算装置がアクセス可能な記憶装置とを有し、

前記演算装置は、請求項 7 から 11 のいずれか一つに記載のデータ生成方法によって生成された疑似データ及び前記訓練データを使用して、前記訓練データ集合に含まれないデータの入力から出力を予測する予測部における学習を実行することを特徴とする学習方法。

【請求項 1 3】

請求項 1 2 に記載の学習方法であって、

前記訓練データを入力したときと前記疑似データを入力したときの内部状態の差、又は

前記訓練データから生成した二つの疑似データの内部状態の差、が小さくなることを良しとする目的関数を追加することを特徴とする学習方法。

【発明の詳細な説明】**【技術分野】****【0001】**

本発明は、機械学習に用いるデータを生成するデータ生成装置に関する。

【背景技術】**【0002】**

昨今、機械学習を応用したシステムやサービスが普及しつつある。例えば、多くの企業は、装置、設備、車両など多種多様な物理的なアセットを運用することによって経済活動を行っている。これらのアセットは電気的又は機械的な理由で不調や故障となることがある。そのため、適宜、予測的又は対処的な処置、すなわち部品や消耗品の交換、ユニットの交換や修繕などによって運用に大きな支障が生じないようすることが重要である。しかし、アセットの複雑化や、熟練した保守員や修理員の不足などから適切な処置が成されない場合がある。そこで、過去の稼働実績や修理履歴などの情報に基づいて適切な処置を推薦することで、保守員や作業員の作業を補い、アセットの安定的な運用を可能とするシステムが検討されている。

【0003】

前述したような機械学習を応用したシステム・サービスにおいて、予測器は、入力及び出力に関わる訓練データ集合から、その入出力の関係性を学習する教師あり学習又は半教師有り学習と称される枠組みに基づいて構築される。この予測器には、訓練データ集合に含まれないデータに対する高い予測性能（汎化性能）が求められる。そこで、昨今ニューラルネットワークをはじめとする予測器の様々なモデルが提案されている。

【0004】

一方、別のアプローチとして、訓練データ集合を適切に変形や変換をして得られる疑似データ集合をもとの訓練データ集合と合わせて学習に用いることで汎化性能向上を目指す方法がある。このような疑似データ生成をデータ拡張と称する。例えば、一般的には、画像であれば、訓練データ集合に対して回転、拡大、縮小、移動などの変形によりデータを拡張するといい。ところが、前述した稼働実績や修理履歴など産業上で取り扱われる多くのデータは、画像のような有効に機能するヒューリスティックなデータ拡張方法が不明な場合も多い。そこで、パラメトリックな分布に従う標本とともに訓練データ集合とを混合することによってデータを拡張する方法が採用される。例えば、小さな標準偏差の正規分布に従う標本の元をもとのデータの元に加えることでデータ拡張できる。ところが、データ拡張された訓練データ集合の分布がもとの訓練データ集合の分布と大きく異なる場合、性能は低下する場合がある。

【0005】

本技術分野の背景技術として、以下の先行技術がある。特許文献 1（特開 2006-343124 号）には、センサ応答から化学物質濃度を推定する技術として、「化学データの補間誤差を確率変数として捉え、補間誤差の確率密度関数を推定する。補間誤差の推定確率密度関数に従う擬似乱数ベクトルを生成し、補間曲面上のベクトルからランダムに選び

10

20

30

40

50

出したベクトルに、擬似補間誤差ベクトルを足し合わせて新たなデータベクトルを生成することを多数繰り返すことで、補間曲面・補間誤差の特徴を反映した多数のデータベクトルである擬似データを生成する。擬似データをニューラルネットワークに学習させる。未知の被検試料についてセンサを適用し、センサ応答を測定する。学習済みとなったニューラルネットワークに、センサ応答を入力し、ニューラルネットワークの出力より複数の化学物質の未知濃度を推定する。」技術が開示されている。

【先行技術文献】

【特許文献】

【0006】

【文献】特開2006-343124号公報

10

【発明の概要】

【発明が解決しようとする課題】

【0007】

しかしながら、特許文献1に記載された技術では、出力データ集合に対する入力データ集合の回帰モデルに対して、誤差に関する分布をカーネル密度推定法で推定し、推定された誤差分布に従う標本の元を推定量に加える操作を行うので、単純に入力データ集合の元に正規分布から得られる集合の元を加える方法に比べ複雑なデータ拡張が達成されるものの、もとの入力データ集合の分布と大きく異なる疑似データ集合が生成されることがある。特に、出力データに対して入力データが1対1となる箇所（単峰的な箇所）と1対多となる箇所（多峰的な箇所）が存在する場合、前述した技術では同一の分布に基づいて変形するため、1対1の箇所では小さな変形を加えるべき箇所で比較的大きな変形がなされ、1対多の箇所では大きな変形を加えるべき箇所で比較的小さな変形がなされ、もとの分布と大きく異なる疑似データ集合となる可能性がある。また、カーネル密度推定法は、訓練データに対して種々のカーネルとカーネルパラメータ（ガウシアンカーネルであればバンド幅）を選択する必要があるなど選択すべき要素が多いという問題がある。

20

【0008】

本発明は、上記を鑑みてなされたものであって、もとの分布と大きく異ならず、訓練データとは異なる疑似データ集合を生成する手段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】

30

本願において開示される発明の代表的な一例を示せば以下の通りである。すなわち、データ集合を生成するデータ生成装置であって、訓練データ集合の各元の入力及び前記訓練データ集合に関する情報の少なくとも一方に基づいて、前記元を変形するための摂動集合を生成する摂動生成部と、前記訓練データ集合及び前記摂動集合から、前記訓練データ集合と異なる新たな疑似データ集合を生成する疑似データ合成部と、前記訓練データ集合と前記疑似データ集合との分布間距離又はそれに関する推定量と、前記摂動集合から得られる訓練データに対する疑似データの摂動の大きさとを算出する評価部と、前記訓練データ集合と前記疑似データ集合との分布間距離を近づけ、摂動の大きさ又は期待値が予め定めた目標値となるように、前記摂動生成部が前記摂動集合の生成に使用するパラメータを更新するパラメータ更新部とを備えることを特徴とする。

40

【発明の効果】

【0010】

本発明の一態様によれば、分布間距離と摂動の大きさとのバランスが取れた、目標摂動量以上に訓練データの分布と異ならない疑似データを生成できる。前述した以外の課題、構成及び効果は、以下の実施例の説明によって明らかにされる。

【図面の簡単な説明】

【0011】

【図1】本実施例のリコメンドシステムの構成を示す図である。

【図2】本実施例のデータ生成・予測器学習部の操作を示す図である。

【図3】本実施例のリコメンドシステムを構成する計算機のハードウェア構成を示す図で

50

ある。

【図4】本実施例の実績データの一例を示す図である。

【図5】本実施例の修理作業データの一例を示す図である。

【図6】本実施例の訓練データ集合の一例を示す図である。

【図7】本実施例におけるモデリングフェーズの処理のフローチャートである。

【図8】本実施例におけるモデリングフェーズの学習処理のフローチャートである。

【図9】本実施例におけるリコメンド処理のフローチャートである。

【図10】本実施例の訓練データ選択画面を示す図である。

【図11】本実施例の疑似データ確認画面を示す図である。

【発明を実施するための形態】

【0012】

以下、適宜図面を参照しながら本発明を実施するための代表的な形態を説明する。

【0013】

<概略>

本発明は、データに基づく機械学習装置に関し、特に、与えられたデータに基づいて別の疑似データを生成し、それを活用して高い汎化性能を持つ予測器を学習する装置に関するものである。本実施例では、前述したアセットが不調又は故障となった場合、アセットの稼働実績、修理履歴などの情報に基づいて適切な処置を推薦するリコメンドシステムで用いる予測器の学習に関するデータ生成・予測器学習装置の概略を説明する。

【0014】

まず、図1を参照してリコメンドシステム全体の処理の流れを説明する。はじめに、稼働実績、修理履歴の収集から予測器学習までの流れ（モデリングフェーズと称する）を説明する。リコメンドシステム11は、アセット13から、アセット13を介してオペレータ16から、及び修理員端末14を介して修理員17から、稼働実績と不具合状況と修理履歴などを収集し、収集した情報を結合した実績データを収集する。ここで、実績データとは、例えばアセット13の稼働時間、アセット13に取り付けられたセンサからの情報、オペレータ16が入力した不具合状況（例えば、異音の発生）、アセット13に実施された修理作業の情報などである。次に、管理者15は、管理端末12を介して、リコメンドシステム11が収集した実績データのうち、データ生成及び予測器の学習に用いるデータを選択する。リコメンドシステム11は、その選択に従ってデータを抽出し、抽出したデータを訓練データとしてデータ生成・予測器学習装置10に送る。データ生成・予測器学習装置10は受信した訓練データを用いてデータを生成し、学習済みモデルを作成する。そして、データ生成・予測器学習装置10は、学習したモデル（学習済みモデル）をリコメンドシステムに返す。

【0015】

次に、修理内容のリコメンドの流れ（リコメンドフェーズと称する）を説明する。リコメンドシステム11は、アセット13から、アセット13を介してオペレータ16から、及び修理員端末14を介して修理員17から、修理作業の情報を除く実績データを収集する。次に、リコメンドシステム11は、前記の学習済みモデルと、前記修理作業の情報を除く実績データから、推奨される1又は複数の修理作業を算出する。そして、その結果を修理員端末14を介して修理員17に提示する。

【0016】

次に、データ生成・予測器学習装置10の処理の概略について説明する。データ生成・予測器学習装置10は、訓練データを受信し、学習済みモデルを作成する。その過程で、高い汎化性能を持った予測器を構築するために、データ生成、データ評価、予測器の三つの構成要素を深層学習（Deep Learning）の一種であるGAN（Generative Adversarial Networks）の枠組みに基づいて学習させる。その際、一般的なGANでは直接疑似データを生成するが、本実施例では、いったん撮動を生成して、生成された撮動を元の訓練データに加えることによって疑似データを生成する。

【0017】

10

20

30

40

50

この結果、摂動に対する目的関数を追加して学習し、学習済みモデルを作成できる。特に、本実施例では、ミニバッチ学習を前提として、そのミニバッチ内の摂動の総和が一定となるようにデータ生成に関する制約を与える。これによって、疑似データが訓練データに分布間距離の観点で近づくことと、疑似データを訓練データから変形させることのトレードオフができる。この結果、疑似データは正規分布等で摂動させた場合と異なり、例えば少しでも動かすと訓練データとしてはあり得ない変数にならず、また訓練データの元をほとんど変形させず、このためデータ拡張による性能劣化を抑制できる。その際、データ拡張の度合いは前述の定数を変えることによって制御できる。

【0018】

一方、予測器の素朴な学習方法は、訓練データに疑似データを混ぜたデータを新たな訓練データセットとして学習させることである。これに加え、訓練データのある元に対して摂動を与えた疑似データが得られるので、それらをラベル無しデータと見做せば半教師有り学習の各種手法が適用できる。例えば、ニューラルネットワークに入力したときの中間層の出力を一致させる処理（本稿では Improved Techniques for Training GANs での表現を参考にしてフィーチャーマッチングと称する）を加えることで、より汎化性能が高い予測器を得ることができる。

10

【0019】

また、前述のフィーチャーマッチングを用いる、データ評価と予測器の一部もしくは全部のニューラルネットワークを共用する、Triple GANなどの方法で予測器を GAN の敵対的学习に参加させる、などの方法によって、ラベル無しの訓練データを有効利用できる。なお、本実施例では一貫して GAN を用いたデータ生成を前提に説明するが、他の方法を用いてもよい。

20

【0020】

<システム構成>

図 1 を参照して本実施例のシステム構成を説明する。本実施例のシステムは、データ生成・予測器学習装置 10 と、リコマンドシステム 11 と、管理者 15 が操作する管理端末 12 と、オペレータ 16 が操作するアセット 13 と、修理員 17 が操作する修理員端末 14 とを有する。これらのシステムの構成要素は、ネットワーク 18 で相互に接続される。ネットワーク 18 自体は LAN (Local Area Network) や WAN (Wide Area Network) などで構成できる。なお、前述したシステム構成は一例であって、構成要素は図示したものに限らない。例えば、データ生成・予測器学習装置 10 とリコマンドシステム 11 が一つの装置として構成されてもよいし、分散処理のため予測器学習装置 10 を複数に分けて構成してもよい。

30

【0021】

次に、図 2 を参照してデータ生成・予測器学習部 101 の詳細について説明する。データ生成・予測器学習部 101 は、摂動生成部 1011 と、疑似データ合成部 1012 と、評価部 1013 と、予測部 1014 と、パラメータ更新部 1015 を備える。

40

【0022】

なお、データ生成・予測器学習部 101 のうち、摂動生成部 1011 と、疑似データ合成部 1012 と、評価部 1013 と、パラメータ更新部 1015 とでデータ生成装置が構成され、予測部 1014 と、パラメータ更新部 1015 とで予測器学習装置が構成される。

【0023】

<機能とハードウェア>

次に図 1 と図 3 を参照して、機能とハードウェアの対応を説明する。

【0024】

データ生成・予測器学習装置 10 が有するデータ生成・予測器学習部 101 と、前処理部 102 と、学習用データ管理部 103 は、CPU (Central Processing Unit) 1H101 が、ROM (Read Only Memory) 1H102 又は外部記憶装置 1H104 に格納されたプログラムを RAM (Read Access Memory) 1H103 に読み込み、通信 I/F (Interface) 1H105、マウスやキーボードなどの入力装置 1H106、ディスプレ

50

イなどの出力装置 1 H 1 0 7 を制御することによって実現される。

【 0 0 2 5 】

リコメンドシステム 1 1 が有するリコメンド部 1 1 1 と、データ管理部 1 1 2 と、集配部 1 1 3 は、C P U (Central Processing Unit) 1 H 1 0 1 が、R O M (Read Only Memory) 1 H 1 0 2 又は外部記憶装置 1 H 1 0 4 に格納されたプログラムをR A M (Read Access Memory) 1 H 1 0 3 に読み込み、通信 I / F (Interface) 1 H 1 0 5 、マウスやキーボードなどの入力装置 1 H 1 0 6 、ディスプレイなどの出力装置 1 H 1 0 7 を制御することによって実現される。

【 0 0 2 6 】

管理端末 1 2 が有する操作部 1 2 1 は、C P U (Central Processing Unit) 1 H 1 0 1 が、R O M (Read Only Memory) 1 H 1 0 2 もしくは外部記憶装置 1 H 1 0 4 に格納されたプログラムをR A M (Read Access Memory) 1 H 1 0 3 に読み込み、通信 I / F (Interface) 1 H 1 0 5 、マウスやキーボードなどの入力装置 1 H 1 0 6 、ディスプレイなどの出力装置 1 H 1 0 7 を制御することで実現される。 10

【 0 0 2 7 】

C P U 1 H 1 0 1 が実行する処理の一部又は全部をハードウェアで構成される演算装置 (A S I C 、F P G A 等) で実行してもよい。

【 0 0 2 8 】

C P U 1 H 1 0 1 が実行するプログラムは、リムーバブルメディア (C D - R O M 、フラッシュメモリなど) 又はネットワークを介してデータ生成・予測器学習装置 1 0 、リコメンドシステム 1 1 及び管理端末 1 2 に提供され、非一時的記憶媒体である不揮発性記憶装置に格納される。このため、計算機システムは、リムーバブルメディアからデータを読み込むインターフェースを有するとよい。 20

【 0 0 2 9 】

データ生成・予測器学習装置 1 0 、リコメンドシステム 1 1 及び管理端末 1 2 の各々は、物理的に一つの計算機上で、又は、論理的又は物理的に構成された複数の計算機上で構成される計算機システムであり、複数の物理的計算機資源上に構築された仮想計算機上で動作してもよい。

【 0 0 3 0 】

<データ構造>

次に、図 4 を参照して、リコメンドシステム 1 1 のデータ管理部 1 1 2 が管理する実績データ 1 D 1 を説明する。実績データ 1 D 1 は、アセット 1 3 や、アセット 1 3 を介してオペレータ 1 6 や、修理員端末 1 4 を介して修理員 1 7 から稼働実績・不具合状況・修理履歴などを収集し、収集したデータを結合して、アセットの修理毎に纏めたデータである。実績データ 1 D 1 は、修理単位を特定するための修理 I D 1 D 1 0 1 と、修理を実施した日時 1 D 1 0 2 と、アセットの設置やオーバーホールからの稼働時間 1 D 1 0 3 と、稼働時の平均温度 1 D 1 0 4 と、稼働時の振動レベル 1 D 1 0 5 と、不具合状況 1 D 1 0 6 と、実施した修理作業を特定するための修理作業 I D 1 D 1 0 7 とを含む。修理作業 I D は、後述するが、実施した作業内容や交換部品などと関連付けられる。 30

【 0 0 3 1 】

なお、本実施例では実績データ 1 D 1 は、前述した項目を含むが、アセットに関する他のデータを含んでもよいし、前述した項目のうち一部の項目を含むものでもよい。

【 0 0 3 2 】

次に、図 5 を参照して、リコメンドシステム 1 1 のデータ管理部 1 1 2 が管理する修理作業データ 1 D 2 を説明する。修理作業データ 1 D 2 は、修理作業を特定する修理作業 I D 1 D 2 0 1 と、作業内容 1 D 2 0 2 と、交換部品 1 D 2 0 3 ~ 1 D 2 0 5 を含む。なお、図 5 に示す例では、交換部品は最大三つまで記録しているが、交換部品の記録数は 3 より大きくて小さくてもよい。また、修理作業データ 1 D 2 は、作業内容及び交換部品以外に、修理作業に関する情報、例えば、使用する工具や消耗品などの情報を含んでもよい。

【 0 0 3 3 】

10

20

30

40

50

次に、図 6 を参照して、データ生成・予測器学習装置 10 の学習用データ管理部 103 が管理する訓練データ集合 1D3 を説明する。訓練データ集合 1D3 は、管理者 15 の指定に基づいて選定された実績データ 1D1 の日時 1D102 や稼働時間 1D103 などに前処理部 102 が前処理を施したデータであって、データを識別するための番号 1D301 と、実績データが数値化された予測器の入力となる入力 1 から入力 1000 (1D302 - 1 ~ 1D302 - 1000) と、修理作業 ID に対応し、予測器の出力となる出力 y1 D303 を含む。なお、本実施例では入力は 1000 個としたが、入力データの数は 1000 よりも多くても少なくてもよい。

【0034】

<処理フロー>

次に、図 7、図 8 を参照して、本実施例におけるモデリングフェーズの処理を説明する。

10

【0035】

図 7 を参照して、全体の流れを説明する。まず、リコマンドシステム 11 の集配部 113 が、アセット 13 及び修理員端末 14 から実績データ 1D1 を収集し、データ管理部 112 に蓄積する (ステップ 1F101)。

【0036】

次に、管理端末 12 の操作部 121 が、実績データ 1D1 からデータ生成及び予測器学習に用いるデータの条件 (期間) と摂動パラメータ探索範囲を管理者 15 から受け付ける。そして、集配部 113 が、受け付けた探索条件に従って、データ管理部 112 から条件に合う実績データ 1D1 を選択し、摂動パラメータ探索範囲と合わせてデータ生成・予測器学習装置 10 の学習用データ管理部 103 に格納する (ステップ 1F102)。なお、摂動パラメータ探索範囲とは、後述する式 (5) の範囲である。

20

【0037】

次に、データ生成・予測器学習装置 10 の前処理部 102 が、学習用データ管理部 103 に格納された選択済みの実績データ 1D1 に文字列やカテゴリカル変数の数値化、量的変数の標準化や正規化等の前処理を施し、訓練データ集合 1D3 を生成、学習用データ管理部 103 に格納する (ステップ 1F103)。

【0038】

次に、データ生成・予測器学習装置 10 のデータ生成・予測器学習部 101 が、訓練データ集合 1D3 に基づいて、データ生成及び予測に関する学習処理を実行し、作成されたモデル (学習済みモデルと称する) を学習用データ管理部 103 に格納する (ステップ 1F104)。なお、学習処理は図 8 を参照して詳述する。

30

【0039】

次に、データ生成・予測器学習装置 10 の学習用データ管理部 103 が、作成されたモデルをリコマンドシステム 11 のデータ管理部 112 に配信 (複製を格納) する (ステップ 1F105)。

【0040】

最後に、管理端末 12 の操作部 121 が、学習済みモデルによって生成される疑似データ集合や、訓練データ集合と疑似データ集合の分布間距離などを管理者 15 に提示し、処理を終了する。管理者 15 は、このような提示情報によって、後述する学習パラメータを変更したり、新しく学習した学習済みモデルを採用するか、従来のモデルを継続利用するかを判断できる。

40

【0041】

次に、図 8 を参照して、本実施例における学習処理について詳しく説明する。なお、本実施例では GAN の一つである Wasserstein GAN (Generative Adversarial Networks) を用いた形態を示す。まず、記号とそれに関する条件を説明する。なお、Wasserstein GAN ではなく、Triple GAN などのほかの GAN の方法を用いてもよいし、分布間距離やそれに関する量に MMD (Maximum Mean Discrepancy) などの他の方法を用いてもよい。また、実際には指定された摂動パラメータ探索範囲についての指定範囲を 10 分割してリニアサーチする等で網羅的に探索し、最も汎化性

50

能が高い学習済みモデルを最終的な学習済みモデルとして選択するとよいが、以下では簡単のため、 θ が 0.2 であるときの処理の流れを説明する。なお、後述する他のパラメータも、 θ と同様に探索してもよい。

【0042】

訓練データ集合 1D3 の入力に関する集合を X と記し、その集合の元 x が倣う分布を P_r と記す。また、疑似データ集合を X_g と記し、その集合の元 x_g が倣う分布を P_g と記す。 P_r と P_g の Wasserstein 距離を $W(P_r, P_g)$ と記す。このとき、 $W(P_r, P_g)$ は、式(1)で表される。

【0043】

【数1】

$$W(P_r, P_g) = \sup_{\|f_w\| \leq 1} E_{x \sim P_r} [f_w(x)] - E_{x \sim P_g} [f_w(x)]$$

10

【0044】

式(1)において、 $\|f_w\| \leq 1$ は、関数 f_w がリブシツ連続であることを表す。また、 $E[\cdot]$ は、期待値を表す。また、関数 f_w はニューラルネットワークで構成され、 w はそのニューラルネットワークのパラメータである。

【0045】

20

x_g は x に摂動 Δx を加えたものであって、以下を満たす。

【0046】

【数2】

$$x_g = x + \Delta x = x + g_\theta(x, z)$$

【0047】

この摂動 Δx は、 x とノイズ z の条件付き確率分布 $P_p(\Delta x | x, z)$ に従う。ここで、ノイズ z は、正規分布や一様分布に従うものとする。また、 g_θ はある x と z から P_p に従う摂動 Δx を生成する関数である。なお、関数 g_θ はニューラルネットワークで構成され、 θ はそのニューラルネットワークのパラメータである。

30

【0048】

次に、入力 x に対して出力の推定値 y_p を算出する関数を $h_\theta(x)$ と記す。なお、関数 h_θ はニューラルネットワークで構成され、 θ はそのニューラルネットワークのパラメータである。前述した記号を使用して処理を説明する。

【0049】

まず、データ生成・予測器学習部 101 の摂動生成部 1011 は、訓練データ集合からサンプリングした部分集合 $X = \{x_m : m = 1, 2, 3, \dots, M\}$ (ミニバッチ集合、本実施例では $M = 100$ とするがほかの値でもかまわない) を抽出し、正規分布からサイズ M の集合 Z のサンプリングし、それらに関数 g_θ を適用し、摂動集合 X_g を生成する (ステップ 1F201)。

40

【0050】

次に、疑似データ合成部 1012 は、前記 X と X_g の各元の和を取ることで、疑似データ集合 $X_g = \{x_g : m = 1, 2, 3, \dots, M\}$ を生成する (ステップ 1F202)。

【0051】

次に、評価部 1013 は、前記 X_g に関数 f_w を適用し、次式によって、分布間距離の一種である Wasserstein 距離の推定量 Wasserstein~ を評価データの一つとして得る (ステップ 1F203)。

【0052】

50

【数3】

$$Wasserstein^{\sim} = \frac{1}{M} \sum_{m=1}^M f_w(x_m) - \frac{1}{M} \sum_{m=1}^M f_w(xg_m)$$

【0053】

次に、データ生成・予測器学習部101の予測部1014は、前記 X と Xg に関数 h を適用し、 X に関する予測データ集合 $Y' = \{y_m, c : m = 1, 2, 3, \dots, M\}$ と、 Xg に関する予測出力 $Y'g = \{y'g_m, c : m = 1, 2, 3, \dots, M\}$ を生成する（ステップ1F204）。ここで、 c はクラスのインデックスを表し、本実施例では、修理作業IDに対応する。

【0054】

次に、データ生成・予測器学習部101のパラメータ更新部1015は、数式（3）で表される推定量 $Wasserstein^{\sim}$ を最大化する方向にパラメータ w を逆誤差伝搬法で更新する。同様に、数式（4）で表される関数 $CrossEntropyLoss$ を最小化する方向にパラメータ w を逆誤差伝搬法で更新する（ステップ1F205）。数式（4）の第一項及び第二項は交差エントロピーを示す。ここで、 y_m, c は、前述と同様に、 X に対応する訓練データの出力データ集合 $Y = \{y_m, c : m = 1, 2, 3, \dots, M\}$ の元であり、インデックスは $y'_m, c, y'g_m, c$ と同じである。また、 u は訓練データ集合由来のパラメータ更新と疑似データ集合由来のパラメータ更新とのバランスを調整するパラメータであり、本実施例では0.5とするが他の値でもよい。なお、数式（4）の第三項は摂動を与えたネットワークの内部状態（中間層の出力）を近づけるような制約を与える。ここで、 u^p_m, c 及び ug^p_m, c は、それぞれ訓練データ集合と疑似データ集合の入力に対する最終層（出力層）の直前の中間層の出力である。 u は、その制約の影響を調整するパラメータであって、本実施例では0.5とするが、他の値でもよい。第三項によって、単にデータ拡張したデータを使って学習することに比べて高い汎化性能を持ったモデルを獲得できる。なお、このステップにおける逆誤差伝搬法を実行する際、摂動生成部101のパラメータ w が更新されないようにするとよい。

【0055】

【数4】

CrossEntropyLoss

$$= - \sum_{m=1}^M \sum_{c=1}^C y_{m, c} \log(y'_{m, c}) - \alpha \sum_{m=1}^M \sum_{c=1}^C y_{m, c} \log(y'g'_{m, c}) \\ + \beta (u^p_{m, c} - ug^p_{m, c})^2$$

【0056】

次に、データ生成・予測器学習部101の摂動生成部1011は、ステップ1F201と同じ手順で摂動集合を生成する（ステップ1F206）。

【0057】

次に、データ生成・予測器学習部101の疑似データ合成部1012は、ステップ1F202と同じ手順で疑似データ集合を生成する（ステップ1F207）。

【0058】

次に、データ生成・予測器学習部101の評価部1013は、前記 Xg に関数 f_w を適用

10

20

30

40

50

し、数式(5)によって関数 g に関するロス Adversarial を評価データの別の一つとして得る(ステップ 1 F 2 0 8)。ここで、 $g(x_m, z) = x_m = x g_m - x_m$ である。また、数式(5)の第一項は、通常の Wasserstein GAN のジェネレータのロス関数が有する項であり、疑似データ集合と訓練データ集合との分布間距離を近づけようとする。一方、第二項は、本発明で採用された項であり、ミニバッチ内の摂動の大きさ(絶対値の総和)が一定値 $\gamma \cdot M$ となるように制約を与える。すなわち、摂動の大きさの期待値を制約している。この結果、訓練データと疑似データとに差異が生じる。これら二つの項の作用によって、本発明の目的である、元の分布と大きく異なることがないが入力データとは異なる疑似データ集合を生成できる。このような疑似データ集合が、元の分布と全く異なることはなく、データ拡張による汎化性能の低下を抑制でき、元となつたデータのラベルを利用できる等、使い勝手がよい疑似データを生成できる。なお、によって、最終的にどのくらい元の訓練データと大きく異なる疑似データを生成するのかを制御できる。本実施例では、 γ は 1.0 とするが、他の値でもよい。なお、前述のとおり、 γ は 0.2 としている。また、摂動の大きさとして絶対値の総和を用いたが、L2 ノルムなどほかの大きさの指標を用いてもよい。

【0059】

【数5】

$$GeneratorLoss = -\frac{1}{M} \sum_{m=1}^M f_w(x g_m) + \lambda \cdot \left(\left(\sum_{m=1}^M |g_\theta(x_m, z)| \right) - \gamma \cdot M \right)^2$$

10

20

【0060】

次に、データ生成・予測器学習部 101 のパラメータ更新部 1015 は、数式(5)で表される Generator Loss を最小化する方向にパラメータ θ を逆誤差伝搬法で更新する(ステップ 1 F 2 0 9)。

【0061】

次に、データ生成・予測器学習部 101 のパラメータ更新部 1015 は、終了条件を満たしているかを確認する。本実施例では、パラメータを所定回数(例えば 10000 回)更新したとき終了条件を満たすとする。終了条件を満たさない場合、ステップ 1 F 2 0 1 に戻り、処理を続ける。一方、終了条件を満たす場合、モデルの学習の処理を終了する(ステップ 1 F 2 1 0)。なお、終了条件としては、数式(4)で表される、いわゆるロス関数の大きさが減少しなくなったタイミングで終了と判定してもよい。

30

【0062】

また、摂動生成部 1011 は、訓練データ集合の入力に関する部分集合 X と、正規分布からサンプリングした集合 Z を用いて、摂動集合 X を生成するが、訓練データ集合の出力に関する部分集合を入力に加えてもよい。これによって、出力の分布を考慮するため、入力と出力の結合分布として、より妥当な疑似データを生成できる。

【0063】

また、訓練データ集合の入力に関する k 近傍密度推定などの確率密度関数の推定量を入力に加えてもよい。これによって、摂動生成部 1011 の学習を高速化かつ安定化できる。

40

【0064】

また、前述では、特定の分布構造を仮定せずに摂動を生成する方法を説明したが、摂動に特定の分布構造(例えば、摂動集合の事後分布を表す正規分布構造などのパラメトリックな分布の母数)を仮定してもよい。その場合、その分布のパラメータ、例えば平均が 0 の正規分布であれば分散を、データ生成の対象とできる。低密度の部分における摂動によって、予測性能を向上でき、摂動生成部 1011 の学習を高速化かつ安定化できる。

【0065】

また、目標摂動量を小さい値から大きい値に変えた場合、目標摂動量の変化に応じて汎化

50

性能が落ち始める直前で止めるような線形探索によって良好な摂動量を得ることができる。

【0066】

また、本実施例では、疑似データと摂動前のデータとでラベルを共用できるので、二つのデータを予測器に入力したときの中間層の出力を近づけることができ、前記フィーチャーマッチングを活用した学習が可能となる。

【0067】

また、本実施例の訓練データ集合はラベル付きであるが、一部にラベルがないデータが含まれている場合、ラベルがないデータに関してはパラメータ（摂動生成部1011）、パラメータw（評価部1013）を、ラベルがあるデータと同じで手順で学習に利用するとともに、式(4)の第三項についてラベルがあるデータと同じ手順の学習により、パラメータ（予測部1014）を利用することで、半教師有り学習ができる。なお、前述したTripleGANのように、予測器を敵対的学習に参加させるように目的関数を定義して半教師有り学習をしてもよい。

10

【0068】

次に、図9を参照して、リコメンド処理の流れを説明する。

【0069】

まず、リコメンドシステム11の集配部113が、修理前（将来的に修理対象となる）のアセット13に関し、アセット13及び修理員端末14から修理作業IDが未記載（None）の実績データ1D1を収集する（ステップ1F301）。

20

【0070】

次に、リコメンドシステム11のリコメンド部111が、データ生成・予測器学習装置10の前処理部102と同様の前処理を施したうえで、学習済みモデルを使って、修理作業IDの予測値（リコメンドと称する）を生成する（ステップ1F302）。

【0071】

次に、リコメンドシステム11のリコメンド部111及び集配部113がアセット13及び修理員端末14にリコメンドを送信する（ステップ1F203）。

【0072】

最後に、アセット13がオペレータ16にリコメンドを提示し、修理員端末14が修理員17にリコメンドを提示して、処理を終了する（ステップ1F204）。

30

【0073】

以上に説明したように、リコメンドシステム11は、アセット13及び修理員端末14から適宜情報を集め、修理のリコメンドを提示することによって、迅速に不調や故障に対応できる。なお、本実施例ではリコメンドシステム11が能動的にリコメンドを生成し提示する形態を示したが、オペレータ16や修理員17の要求に応じてリコメンドを生成し提示する処理を実行してもよい。

【0074】

<ユーザインターフェース>

次に、図10を参照して、管理者15がデータ生成及び予測器学習に用いる実績データ1D1を選択するために使用する訓練データ選択画面1G1を説明する。訓練データ選択画面1G1は、管理端末12の操作部121に表示される。

40

【0075】

訓練データ選択画面1G1は、期間開始日設定ボックス1G101と、期間終了日設定ボックス1G102と、摂動パラメータ探索範囲下限設定ボックス1G103と、摂動パラメータ探索範囲上限設定ボックス1G104と、設定ボタン1G105を含む。

【0076】

期間開始日設定ボックス1G101で開始日を指定し、期間終了日設定ボックス1G102で終了日を指定することによって、開始日から終了日までの期間の実績データ1D1が訓練データとして選択される。

【0077】

摂動パラメータ探索範囲下限設定ボックス1G103で摂動パラメータ探索範囲の下限を

50

設定し、撮動パラメータ探索範囲上限設定ボックス 1 G 1 0 4 で撮動パラメータ探索範囲の上限を設定することによって、撮動の総量を変えて最良なモデルを学習できる。なお、図示したように撮動パラメータ探索範囲の下限及び上限を設定するのではなく、撮動パラメータを設定するための設定ボックスを設けてもよい。

【 0 0 7 8 】

設定ボタン 1 G 1 0 5 が操作（例えば、クリック）されると、前述した学習に用いる実績データ 1 D 1 の期間と撮動パラメータ探索範囲が、データ生成・予測器学習装置 1 0 の学習用データ管理部 1 0 3 に格納される。

【 0 0 7 9 】

次に、図 1 1 を参照して、管理者 1 5 が学習済みモデルによって生成される疑似データを 10 目視確認するために使用する疑似データ確認画面 1 G 2 を説明する。疑似データ確認画面 1 G 2 は、管理端末 1 2 の操作部 1 2 1 に表示される。

【 0 0 8 0 】

疑似データ確認画面 1 G 2 は、X 軸成分指定リストボックス 1 G 2 0 1 と、Y 軸成分指定リストボックス 1 G 2 0 2 と、比較ビュー 1 G 2 0 3 と、分布間距離ボックス 1 G 2 0 4 を含む。

【 0 0 8 1 】

X 軸成分指定リストボックス 1 G 2 0 1 には、比較ビュー 1 G 2 0 3 の X 軸に割り当てられる前処理済み訓練データ 1 D 3 の入力（例えば入力 1）が設定される。同様に、Y 軸成分指定リストボックス 1 G 2 0 2 には、比較ビュー 1 G 2 0 3 の Y 軸に割り当てられる前処理済み訓練データ 1 D 3 の入力（例えば入力 3）が設定される。この結果、設定された二つの入力に関する前処理済み訓練データ 1 D 3（図中のもとデータ）と、生成された疑似データとが散布図として比較ビュー 1 G 2 0 3 に表示される。管理者 1 5 が比較ビュー 1 G 2 0 3 を見ることによって、入力されるデータがどのように拡張されたかを目視で確認できる。これは、例えば、少数のデータがよくばらついている箇所では、データを追加収集すべきという判断ができる。

【 0 0 8 2 】

一方、分布間距離ボックス 1 G 2 0 4 には、M M D で算出された全入力に関する分布間距離が表示される。これは疑似データがもとの前処理済み訓練データ 1 D 3 と異なる程度を確認するために使うことができる。ここで、評価部 1 0 1 3 の評価結果を用いてもよいが、学習条件によって学習されるW a s s e r s t e i n 距離の推定量が異なるため、本実施例ではM M D を用いる。

【 0 0 8 3 】

以上に説明したように、本発明の実施例によれば、パラメータ更新部 1 0 1 5 は、訓練データ集合と疑似データ集合との分布間距離を近づけ、撮動の大きさ又は期待値が予め定めた目標値となるように、撮動生成部 1 0 1 1 が撮動集合の生成に使用するパラメータを更新するので、与えられた訓練データ集合の各元の特徴を考慮の上、疑似データ全体として訓練データ集合に対して分布間距離もしくはそれに関する推定量が小さくなるような撮動を加えることができ、目標撮動量以上に訓練データの分布と異ならない疑似データを生成できる。

【 0 0 8 4 】

また、撮動生成部 1 0 1 1 は、訓練データ集合の各元の入力又は訓練データ集合に関する情報、及び訓練データ集合の各元の出力又はそれに関する情報に基づいて撮動集合を生成するので、分布間距離と撮動の大きさのトレードオフの観点で、出力の分布が考慮された入力と出力の結合分布として、より妥当な疑似データを生成できる。

【 0 0 8 5 】

また、撮動生成部 1 0 1 1 は、前記訓練データ集合の各元の入力又は前記訓練データ集合に関する情報に加えて、前記訓練データ集合の入力に関する確率密度関数（例えば k 近傍密度推定）の推定量に基づいて撮動集合を生成するので、撮動生成部 1 0 1 1 の学習を高速化かつ安定化できる。

10

20

30

40

50

【0086】

また、撮動生成部1011は、撮動集合の事後分布を表すパラメトリックな分布（例えば正規分布）の母数を生成することによって、撮動集合を生成するので、低密度の部分における撮動によって、予測性能を向上でき、学習を高速化かつ安定化できる。

【0087】

また、撮動生成部1011が使用するパラメータ値又はその範囲を入力可能なインターフェース画面の表示データ（訓練データ選択画面1G1）を生成するので、撮動量を変えて最良なモデルを学習するための条件を付与できる。

【0088】

また、訓練データ集合の各元と疑似データ集合の各元とが表された散布図の表示データを生成するので、入力されるデータがどのように拡張されたかを確認できる。 10

【0089】

また、予測部1014は、前述したデータ生成装置が生成した疑似データ及び訓練データを使用して学習をするので、予測性能を向上でき、学習を高速化かつ安定化できる。

【0090】

また、予測部1014は、ニューラルネットワークで構成され、訓練データを入力したときと疑似データを入力したときの内部状態の差が小さくなることを良しとする目的関数（例えば式（4）の第三項）を追加するので、より高い汎化性能を持ったモデルを獲得できる。なお、ある訓練データから生成した二つの疑似データの内部状態の差が小さくなることを良しとする目的関数でもよい。 20

【0091】

なお、本発明は前述した実施例に限定されるものではなく、添付した特許請求の範囲の趣旨内における様々な変形例及び同等の構成が含まれる。例えば、前述した実施例は本発明を分かりやすく説明するために詳細に説明したものであり、必ずしも説明した全ての構成を備えるものに本発明は限定されない。また、ある実施例の構成の一部を他の実施例の構成に置き換えてよい。また、ある実施例の構成に他の実施例の構成を加えてよい。また、各実施例の構成の一部について、他の構成の追加・削除・置換をしてもよい。

【0092】

また、前述した各構成、機能、処理部、処理手段等は、それらの一部又は全部を、例えば集積回路で設計する等により、ハードウェアで実現してもよく、プロセッサがそれぞれの機能を実現するプログラムを解釈し実行することにより、ソフトウェアで実現してもよい。 30

【0093】

各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリ、ハードディスク、S S D (S o l i d S t a t e D r i v e) 等の記憶装置、又は、I C カード、S D カード、D V D 等の記録媒体に格納することができる。

【0094】

また、制御線や情報線は説明上必要と考えられるものを示しており、実装上必要な全ての制御線や情報線を示しているとは限らない。実際には、ほとんど全ての構成が相互に接続されていると考えてよい。

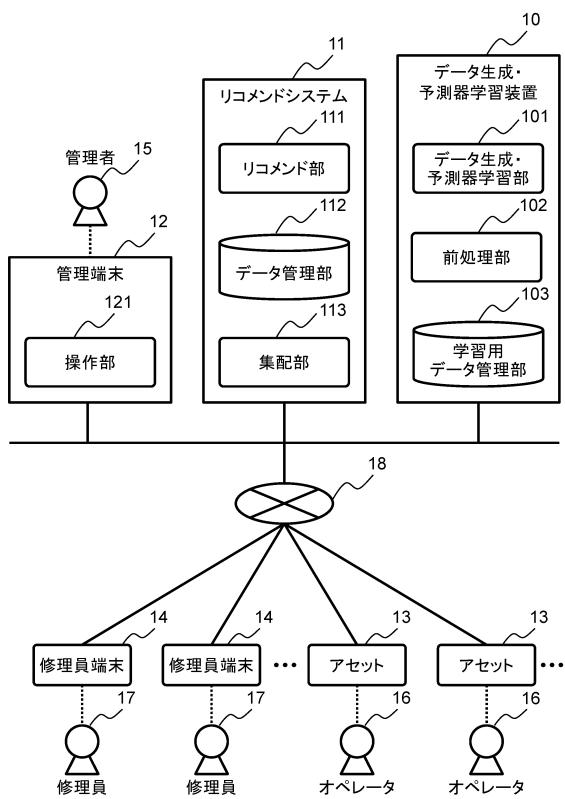
【符号の説明】 40

【0095】

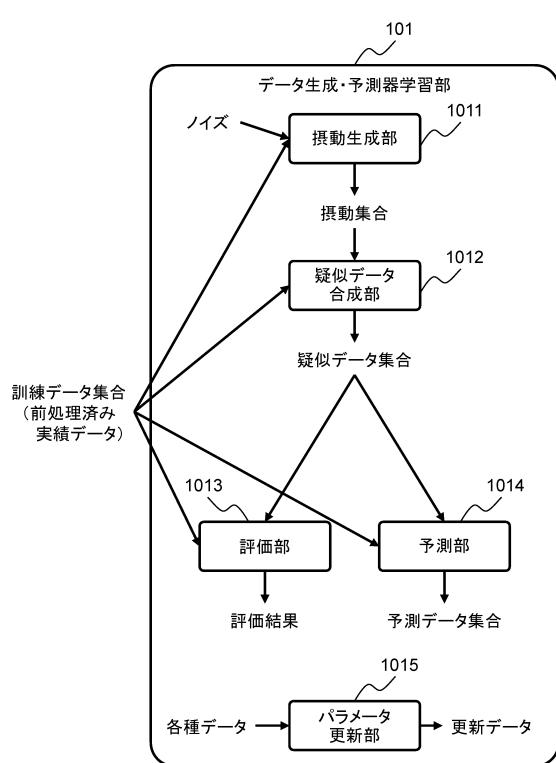
- 10 データ生成・予測器学習装置
- 101 データ生成・予測器学習部
- 102 前処理部
- 1011 摄動生成部
- 1012 疑似データ合成部
- 1013 評価部
- 1014 予測部
- 1015 パラメータ更新部

【図面】

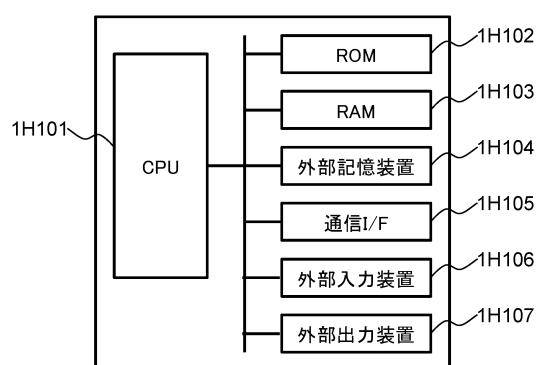
【図 1】



【図 2】



【図 3】



【図 4】

1D101	1D102	1D103	1D104	1D105	1D106	1D107
修理ID	日時	稼働時間	稼働時平均温度	稼働時振動レベル	不具合状況	修理作業ID
1001	2018/1/1 14:32:24	1711	98	Lv1	エンジンがかかりにくい	3
1002	2018/1/1 14:32:26	1204	88	Lv3	変速部の振動が大きい	36
1003	2018/1/1 14:32:	684	78	Lv1	エンジンが突然とまる	41
1004	2018/1/1 14:32:16	2103	89	Lv1	セルモータがまわらない	1
...

10

20

30

40

50

【図 5】

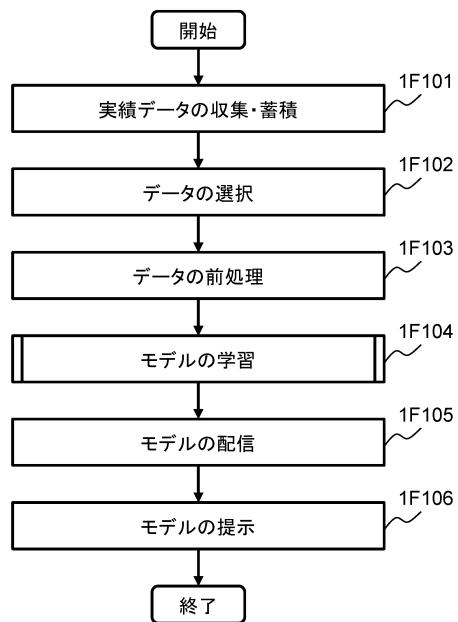
修理作業ID	作業内容	交換部品1	交換部品2	交換部品3
1	エンジンオイルの注油	エンジンオイル	None	None
2	変速部へのグリースアップ	変速用グリス	None	None
3	プラグ交換	スパークプラグ	None	None
...

【図 6】

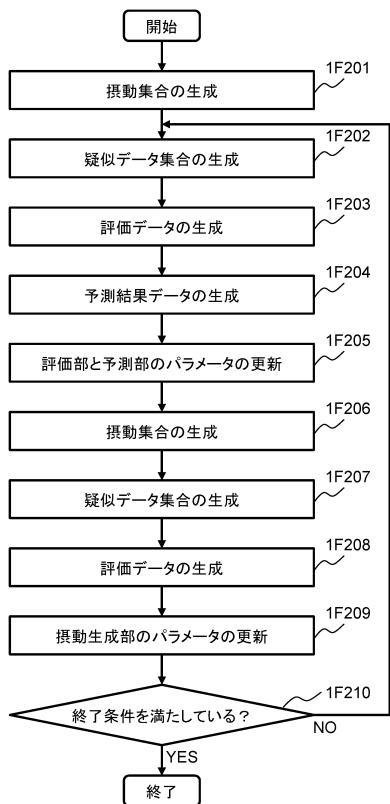
番号	入力1	入力2	入力3	...	入力1000	出力y
0	1.23	1.14	1	...	0	3
1	1.11	1.06	0	...	1	36
2
3
...

10

【図 7】



【図 8】



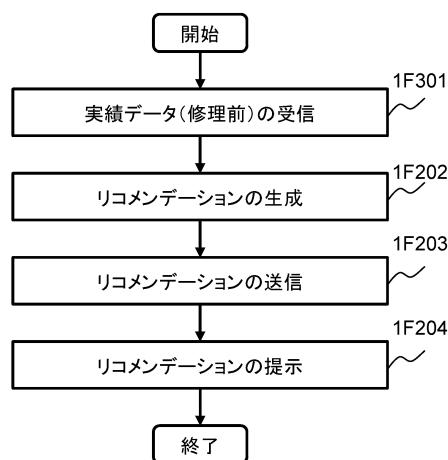
20

30

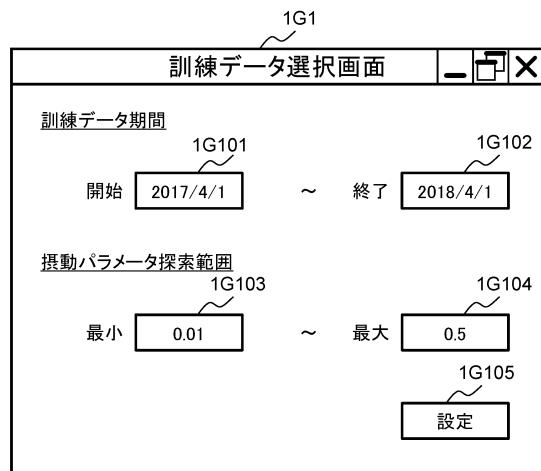
40

50

【図 9】



【図 10】



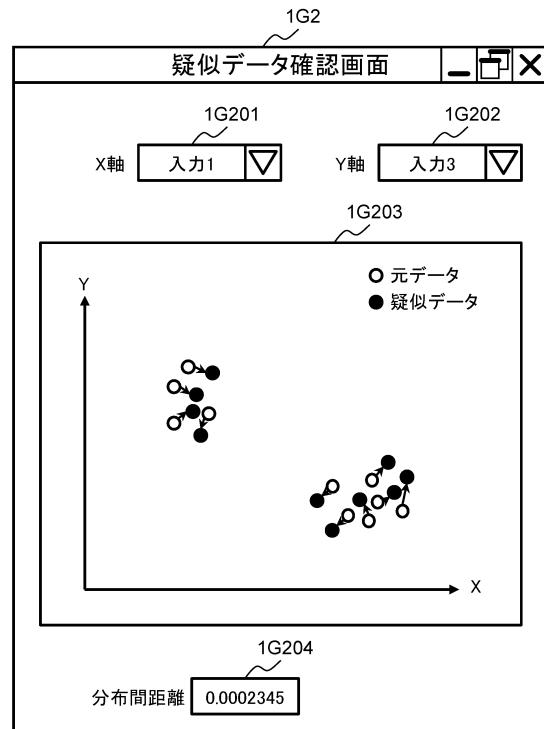
10

20

30

40

【図 11】



50

フロントページの続き

(56)参考文献

LUO, Yun et al. , EEG Data Augmentation for Emotion Recognition Using a Conditional Wasserstein GAN , IEEE , 2018年10月29日 , pp.2535-2538 , 検索日[2020.01.27]、インターネット:<URL:<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8512865>> , DOI:<10.1109/EMBC.2018.8512865> , 第 2 5 3 7 ページ右カラム第 2 0 行-第 2 5 3 8 ページ左カラム第 2 8 行
SALIMANS, Tim et al. , Improved Techniques for Training GANs , v1 , 2016年06月10日 , pp.1-10 , [検索日 2020.01.27]、インターネット:<URL:<https://arxiv.org/pdf/1606.03498.pdf>> , 第 2 ページ第 3 4 - 4 8 行
LOU, Huan et al. , One-dimensional Data Augmentation Using a Wasserstein Generative Adversarial Network with Supervised , IEEE , 2018年07月09日 , pp.1896-1901 , [検索日 20 20.01.27]、インターネット:<URL:<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8407436>> , DOI: 10.1109/CCDC.2018.8407436
VOLPI, R., et al. , Generalizing to Unseen Domains via Adversarial Data Augmentation , Advances in Neural Information Processing Systems 31 (NeurIPS 2018) [online] , 2018年12月03日 , [retrieved on 2022-04-13]. Retrieved from the Internet: URL: <https://proceedings.neurips.cc/paper/2018/hash/1d94108e907bb8311d8802b48fd54b4a-Abstract.html>

(58)調査した分野

(Int.Cl. , D B 名)

G 0 6 N 2 0 / 0 0

G 0 6 N 3 / 0 8