

(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2020-52936
(P2020-52936A)

(43) 公開日 令和2年4月2日(2020.4.2)

(51) Int.Cl.
G06N 20/00 (2019.01)

F I
G06N 99/00 150

テーマコード (参考)

審査請求 有 請求項の数 6 O L (全 14 頁)

(21) 出願番号 特願2018-184242 (P2018-184242)
(22) 出願日 平成30年9月28日 (2018.9.28)
(11) 特許番号 特許第6622369号 (P6622369)
(45) 特許公報発行日 令和1年12月18日 (2019.12.18)

(71) 出願人 394013002
三菱電機インフォメーションシステムズ株式会社
東京都港区芝浦四丁目13番23号
(74) 代理人 100110423
弁理士 曾我 道治
(74) 代理人 100166235
弁理士 大井 一郎
(72) 発明者 小池 賢一
東京都港区芝浦四丁目13番23号 三菱電機インフォメーションシステムズ株式会社内

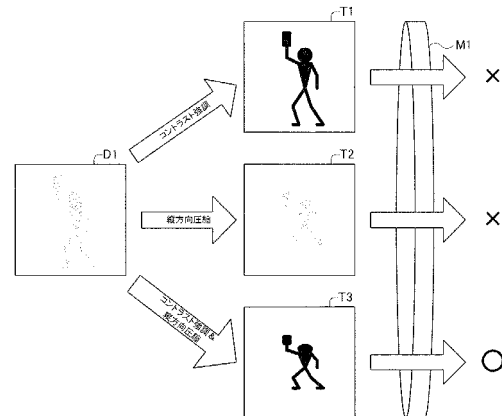
(54) 【発明の名称】 訓練データを生成する方法、コンピュータおよびプログラム

(57) 【要約】

【課題】 質の良い訓練データを自動的に生成することができる方法等を提供する。

【解決手段】 データ生成装置10は、検証データD1～D3を用いて機械学習を行うことにより第1学習済みモデルM1を生成し、第1学習済みモデルM1を用いて検証データD1～D3を分類し、正しく分類されなかった検証データのうちから基準データD1を選択し、基準データD1から変換済データT1～T3を生成し、変換済データT1～T3を分類し、正しく分類された変換データT3に係る変換の逆変換を訓練データK1, K2に対して施すことにより、新たな訓練データN2, N3を生成する。

【選択図】 図4



【特許請求の範囲】**【請求項 1】**

機械学習に用いられる学習データに含まれる訓練データを生成する方法であって、
コンピュータが、訓練データを用いて機械学習を行うことにより、第 1 学習済みモデル
を生成するステップと、

コンピュータが、前記第 1 学習済みモデルを用いて、少なくとも 1 つの検証データにつ
いて出力値を取得するステップと、

前記第 1 学習済みモデルによる出力値が不適切であった検証データのうちから、コンピ
ュータが、少なくとも 1 つの基準データを選択するステップと、

コンピュータが、前記基準データに対して複数の異なる変換を施すことにより、複数の
異なる変換済データを生成するステップと、

コンピュータが、前記第 1 学習済みモデルを用いて、各前記変換済データについて出力
値を取得するステップと、

コンピュータが、前記変換済データのうち出力値が適切であったものに施されていた前
記変換の逆変換を、少なくとも 1 つの訓練データに対して施すことにより、少なくとも 1
つの新たな訓練データを生成するステップと、

を備える、方法。

【請求項 2】

機械学習に用いられる学習データに含まれる訓練データを生成する方法であって、

コンピュータが、訓練データを用いて機械学習を行うことにより、第 1 学習済みモデル
を生成するステップと、

コンピュータが、前記第 1 学習済みモデルを用いて、少なくとも 1 つの検証データにつ
いて出力値を取得するステップと、

前記第 1 学習済みモデルによる出力値が適切であった検証データのうちから、コンピ
ュータが、少なくとも 1 つの基準データを選択するステップと、

コンピュータが、前記基準データに対して複数の異なる変換を施すことにより、複数の
異なる変換済データを生成するステップと、

コンピュータが、前記第 1 学習済みモデルを用いて、各前記変換済データについて出力
値を取得するステップと、

コンピュータが、前記変換済データのうち出力値が不適切であったものに施されていた
前記変換を、少なくとも 1 つの訓練データに対して施すことにより、少なくとも 1 つの新
たな訓練データを生成するステップと、

を備える、方法。

【請求項 3】

前記学習データに含まれる訓練データおよび検証データは 2 次元画像データであり、前
記変換は、幾何学的線形変換、色値の変換、ニューラルネットで抽出した特徴量の追加や
削除または、ニューラルネットで抽出した特徴量の強調や弱める変換を含む、請求項 1 ま
たは 2 に記載の方法。

【請求項 4】

前記学習データに含まれる訓練データおよび検証データは音声データであり、前記変換
は、音の高さの変更、音声の再生スピードの変更、雑音の付加、雑音の除去、ローパスフ
ィルタの適用、ハイパスフィルタの適用、ニューラルネットで抽出した特徴量の追加や削
除または、ニューラルネットで抽出した特徴量の強調や弱める変換を含む、請求項 1 また
は 2 に記載の方法。

【請求項 5】

請求項 1 ~ 4 のいずれか一項に記載の方法を実行するコンピュータ。

【請求項 6】

コンピュータに請求項 1 ~ 4 のいずれか一項に記載の方法を実行させるプログラム。

【発明の詳細な説明】**【技術分野】**

10

20

30

40

50

【0001】

本発明は、訓練データを生成する方法等に関する。

【背景技術】

【0002】

機械学習において、訓練データの質のみならず、量もまた重要である。とくに、データを高い精度で分類（クラスタリング）または予測するための学習済みモデルを生成するためには、正解が付与された学習データが大量に必要となる。

【0003】

このような訓練データを準備するためには、大量のデータを人手で分類して正解を付与したり、人手で予測値を準備したりする必要があり、多大な労力が必要となる。このために、訓練データを自動的に生成する方法の例が、特許文献1に記載されている。

10

【先行技術文献】

【特許文献】

【0004】

【特許文献1】特開2014-178229号公報

【発明の概要】

【発明が解決しようとする課題】

【0005】

しかしながら、従来の技術では、質の良い学習データを生成することが困難であるという問題があった。

20

【0006】

たとえば特許文献1の構成では訓練データそのものを生成することができず、元の訓練データとは無関係に特徴量のみに基づいて新たな訓練データを作成している。このため、実データ（たとえば写真）とはかけ離れたデータ（たとえば画像データ形式ではないデータや、人間が写真と認識できないようなデータ）が生成されてしまうおそれがある。同様の問題は、画像データだけでなく、たとえば音声データでも発生する可能性がある。

【0007】

この発明は、このような問題点を解決するためになされたものであり、質の良い訓練データを自動的に生成することができる方法、コンピュータおよびプログラムを提供することを目的とする。

30

【課題を解決するための手段】

【0008】

上述の問題点を解決するため、この発明に係る方法は、機械学習に用いられる学習データに含まれる訓練データを生成する方法であって、

コンピュータが、訓練データを用いて機械学習を行うことにより、第1学習済みモデルを生成するステップと、

コンピュータが、前記第1学習済みモデルを用いて、少なくとも1つの検証データについて出力値を取得するステップと、

前記第1学習済みモデルによる出力値が不適切であった検証データのうちから、コンピュータが、少なくとも1つの基準データを選択するステップと、

40

コンピュータが、前記基準データに対して複数の異なる変換を施すことにより、複数の異なる変換済データを生成するステップと、

コンピュータが、前記第1学習済みモデルを用いて、各前記変換済データについて出力値を取得するステップと、

コンピュータが、前記変換済データのうち出力値が適切であったものに施されていた前記変換の逆変換を、少なくとも1つの訓練データに対して施すことにより、少なくとも1つの新たな訓練データを生成するステップと、
を備える。

また、この発明に係る方法は、機械学習に用いられる学習データに含まれる訓練データを生成する方法であって、

50

コンピュータが、訓練データを用いて機械学習を行うことにより、第1学習済みモデルを生成するステップと、

コンピュータが、前記第1学習済みモデルを用いて、少なくとも1つの検証データについて出力値を取得するステップと、

前記第1学習済みモデルによる出力値が適切であった検証データのうちから、コンピュータが、少なくとも1つの基準データを選択するステップと、

コンピュータが、前記基準データに対して複数の異なる変換を施すことにより、複数の異なる変換済データを生成するステップと、

コンピュータが、前記第1学習済みモデルを用いて、各前記変換済データについて出力値を取得するステップと、

10

コンピュータが、前記変換済データのうち出力値が不適切であったものに施されていた前記変換を、少なくとも1つの訓練データに対して施すことにより、少なくとも1つの新たな訓練データを生成するステップと、
を備える。

特定の態様によれば、前記学習データに含まれる訓練データおよび検証データは2次元画像データであり、前記変換は、幾何学的線形変換または色値の変換、ニューラルネットで抽出した特徴量の追加や削除、ニューラルネットで抽出した特徴量の強調や弱める変換を含む。

特定の態様によれば、前記学習データに含まれる訓練データおよび検証データは音声データであり、前記変換は、音の高さの変更、音声の再生スピードの変更、雑音の付加、雑音の除去、ローパスフィルタの適用、またはハイパスフィルタの適用、ニューラルネットで抽出した特徴量の追加や削除、ニューラルネットで抽出した特徴量の強調や弱める変換を含む。

20

また、この発明に係るコンピュータは、上述の方法を実行する。

また、この発明に係るプログラムは、コンピュータに上述の方法を実行させる。

【発明の効果】

【0009】

この発明に係る方法等によれば、実際の学習データに変換を施すことで新たな学習データを生成するので、より質の良い学習データを生成することができる。

【図面の簡単な説明】

30

【0010】

【図1】本発明の実施の形態1に係るコンピュータの構成の例を示す図である。

【図2】実施の形態1に係る学習データの例を示す図である。

【図3】実施の形態1に係る処理の流れを説明するフローチャートである。

【図4】実施の形態1における基準データに対する変換の例を示す図である。

【図5】実施の形態1において生成される新たな訓練データの例を示す図である。

【図6】実施の形態2に係る検証データの例を示す図である。

【図7】実施の形態2に係る処理の流れを説明するフローチャートである。

【図8】実施の形態2における基準データに対する変換の例を示す図である。

【図9】実施の形態2において生成される新たな訓練データの例を示す図である。

40

【図10】図1のコンピュータの各機能を専用のハードウェアである処理回路で実現する場合を示した構成図である。

【図11】図1のコンピュータの各機能をプロセッサおよびメモリを備えた処理回路により実現する場合を示した構成図である。

【発明を実施するための形態】

【0011】

以下、この発明の実施の形態を添付図面に基づいて説明する。

実施の形態1 .

図1に、本発明の実施の形態1に係るデータ生成装置10の構成の例を示す。データ生成装置10は、機械学習に用いられる学習データを生成する装置(コンピュータ)として

50

機能する。

【0012】

図1に示すように、データ生成装置10は公知のコンピュータとしての構成を有し、演算を行う演算手段11と、情報を格納する記憶手段12とを備える。演算手段11はたとえばCPU(中央処理装置)を含み、記憶手段12はたとえば半導体メモリおよびHDD(ハードディスクドライブ)を含む。

【0013】

記憶手段12は、学習データを記憶する。学習データは、データについて出力値を取得する学習済みモデルを生成するための学習データである。学習済みモデルは、たとえばデータを複数のクラスのいずれかに分類するためのモデルであってもよい。その場合には、学習済みモデルの出力値は、たとえばそのデータがいずれのクラスに属するかを表す値となる。また、学習済みモデルは、未来のデータを予測するためのモデルであってもよい。その場合には、学習済みモデルの出力値は、たとえば未来のデータの予測値となる。以下では、データを分類するためのモデルについて説明するが、本発明はデータの予測を行うモデルにも応用が可能である。機械学習処理における学習データの具体的な利用方法は適宜設計可能であるが、訓練データと検証データを含んでいる。

10

【0014】

学習データは、所定のデータ形式を有するデータ単位の集合である。各データ単位は、分類対象となるデータ部分と、正解ラベルとを含んでいる。なお、本明細書において、「データを分類する」という表現および「データを認識する」という表現は、いずれも、「データの正解ラベルを推定する」という意味である場合がある。以下では、データが画像データである場合について説明するが、本発明はデータが音声データである場合にも応用が可能である。

20

【0015】

図2に、実施の形態1における検証データの例を示す。検証データは検証データD1~D3を含む。本実施形態では、各データ単位において、分類対象となるデータ部分は2次元画像である。具体例として、あるデータ単位のデータ部分は人物の画像を表し、当該データ単位の正解ラベルは「人物」である。この場合には、当該データ単位は[人物]というクラスに属するということができる。図2の検証データD1~D3はこれに該当する。また、別のデータ単位は人物でない物の画像を表し、当該データ単位の正解ラベルは「非人物」である。この場合には、当該データ単位は[非人物]というクラスに属するということができる。3個以上のクラスが定義されてもよい。

30

【0016】

訓練データも、検証データと同一の形式で定義される。この例では、これらの訓練データは、データを「人物」または「非人物」のクラスに分類する学習済みモデルを生成するための訓練データであるということができる。すなわち、これらの訓練データを用いた機械学習により生成される学習済みモデルは、画像が人物を表すものであるか否かを認識するためのモデルとなる。

【0017】

データ生成装置10の記憶手段12はプログラム(図示せず)も格納しており、演算手段11がこのプログラムを実行することによって、データ生成装置10は本明細書に記載される機能を実現する。すなわち、このプログラムは、本明細書に記載される方法を、コンピュータに実行させるものである。

40

【0018】

データ生成装置10は、公知のコンピュータが通常備える他の構成要素を備えてもよい。たとえば、出力装置であるディスプレイおよびプリンタ、入力装置であるキーボードおよびマウス、通信ネットワークに対する入力装置と出力装置とを兼ねるネットワークインタフェース、等を備えてもよい。

【0019】

図3は、実施の形態1においてデータ生成装置10が実行する処理の流れを説明するフ

50

ローチャートである。図3の処理において、まずデータ生成装置10は、訓練データを用いて機械学習を行うことにより、第1学習済みモデルM1を生成する(ステップS1)。第1学習済みモデルM1は、データを複数のクラスのいずれかに分類するモデルである。すなわち、データについて、そのデータがいずれのクラスに属するかを表す値を、出力値として出力する。第1学習済みモデルM1はどのような形式のモデルであってもよいが、たとえばニューラルネットワークを用いて構成することができる。第1学習済みモデルM1の具体的な構造は当業者が適宜設計可能である。

【0020】

次に、データ生成装置10は、第1学習済みモデルM1を用いて、検証データを分類する(ステップS2)。すなわち、検証データについて出力値を取得する。ステップS2は少なくとも1つの検証データ(たとえば検証データD1)に対して行われればよいが、すべての検証データに対して行ってもよい。

10

【0021】

ここでは、図2に示す検証データのうち、検証データD1は正しく分類されず(たとえば人物でないとして認識され)、検証データD2およびD3は正しく分類された(たとえば人物であると認識された)とする。検証データD1の分類を誤った理由としては、たとえば画像中の人物が縦方向に長すぎ、かつ色が薄いためである、という理由が考えられるものとする。

【0022】

次に、データ生成装置10は、第1学習済みモデルM1を用いて正しく分類されなかった検証データ(すなわち第1学習済みモデルM1による出力値が不適切であった検証データ)のうちから、少なくとも1つの基準データを選択する(ステップS3)。ここで、「正しく分類されなかった」とは、たとえば誤ったクラスに分類されたもの(人物の画像であるのに[非人物]のクラスに分類された等)を意味するが、いずれのクラスにも分類されなかったものを含んでもよい。

20

【0023】

選択の基準は適宜設計可能であり、たとえば正しく分類されなかった検証データのうちからランダムに選択するようにしてもよいし、正しく分類されなかった検証データをすべて基準データとして選択するようにしてもよい。ここでは、検証データD1が基準データとして選択されたものとする。

30

【0024】

次に、データ生成装置10は、基準データに対して複数の異なる変換を施すことにより、複数の異なる変換済データを生成する(ステップS4)。

図4に、基準データに対する変換の例を示す。検証データD1に対してコントラストを強調する変換を施すことにより、変換済データT1が生成される。同様に、検証データD1に対して縦方向に圧縮する変換を施すことにより、変換済データT2が生成され、検証データD1に対してコントラストを強調する変換および縦方向への圧縮に圧縮する変換を施すことにより、変換済データT3が生成される。

【0025】

このような変換に対応する具体的な演算処理は、公知の画像処理ソフトウェアまたは他の技術を用いて適宜設計可能である。

40

【0026】

ここで、変換済データT1に係る変換と、変換済データT3に係る変換とは、その一部が共通している(いずれもコントラストを強調する変換を含む)が、全体として異なる変換となっている。このように、変換済データT1~T3に対しては、互いに異なる変換が施されたということが出来る。このように、各変換済データは、基準データに対してそれぞれ異なる変換を施すことにより生成される。

【0027】

次に、データ生成装置10は、第1学習済みモデルM1を用いて、変換済データT1~T3を分類する(ステップS5)。すなわち、第1学習済みモデルM1を用いて、変換済

50

データT1～T3について出力値を取得する。ここでは、変換済データT1およびT2は正しく分類されず（たとえば人物でないと認識され）、変換済データT3は正しく分類された（たとえば人物であると認識された）ものとする。

【0028】

なお、ここですべての変換済データが正しく分類された場合には、処理をステップS3に戻し、基準データの選択から処理をやり直してもよい。

【0029】

ここで、検証データD1の分類は誤っていたにも関わらず、検証データD1から生成された変換済データT3の分類が正しく行われるようになった理由としては、変換済データT3のコントラストおよび縦横比率が、他の多くの訓練データと同程度に変更されていたため、第1学習済みモデルM1が正しく分類できるようになった、という理由が考えられるものとする。これを言い換えると、検証データD1のようにコントラストが薄く、かつ縦方向に長い画像が、学習データ中に不足していたものと考えられる。このため、検証データD1のようにコントラストが薄く、かつ縦方向に長い画像を訓練データ中に追加して新たに機械学習をやり直せば、検証データD1を正しく分類できる学習済みモデルが生成される可能性が高まることが期待される。

10

【0030】

次に、データ生成装置10は、変換の内容に基づき、新たな訓練データを生成する（ステップS6）。より具体的には、変換済データのうち正しく分類されたもの（すなわち出力値が適切であったもの）に施されていた変換の逆変換を、全ての訓練データに施す。この例では、正しく分類された変換済データT3に施されていた変換は「コントラストを強調する変換および縦方向に圧縮する変換」であるので、これに対する逆変換は、たとえば「コントラストを緩和する変換および縦方向に伸長する変換」となる。なお、この例では変換を構成する要素の順序は可換であるが、そうでない場合には各要素の順序を逆転してもよい。

20

【0031】

なお、上述のように新たな学習データの元となる訓練データは、少なくとも1つであればよい。

【0032】

図5に、新たな訓練データの例を示す。訓練データK1に対して「コントラストを緩和する変換および縦方向に伸長する変換」を施すことにより、新たな訓練データN2が生成される。同様に、訓練データK2に対して「コントラストを緩和する変換および縦方向に伸長する変換」を施すことにより、新たな訓練データN3が生成される。同様に、全ての訓練データに対して、変換が施される。

30

【0033】

次に、データ生成装置10は、ステップS1で用いた訓練データと、ステップS6で生成された新たな訓練データとを用いて機械学習を行うことにより、第2学習済みモデルを生成する（ステップS7）。

【0034】

ここで、新たな訓練データN2およびN3は、検証データD1のようにコントラストが薄く、かつ縦方向に長い画像となっているので、これらを用いて生成された第2学習済みモデルは、検証データD1を正しく分類できる可能性が高いといえる。

40

【0035】

以上説明するように、本発明の実施の形態1に係るデータ生成装置10によれば、検証データD1を正しく分類するための質の良い学習データを自動的に生成することができる。

【0036】

ステップS5では、検証データD1に異なる変換を施した変換済データT1～T3を分類することにより、ステップS6で施す変換の内容を決定している。正しく分類されなかった検証データの全てまたは一部を選択し、ステップS5を実行する。ステップS6で施

50

す変換は、ステップ5で正しく分類されるようになったデータが多い変換の逆変換を、施す変換の内容と決定するように構成してもよい。このとき、一番正しく分類されるようになったデータが多い変換の逆変換だけを施してもよいし、上位複数の変換の逆変換を施すように構成してもよい。

ステップS2で検証データD2、D3のように正しく分類されていた検証データが、ステップS5で正しく分類されなくなる場合はその検証データ数をカウントしてもよい。この検証データ数が一定値より多い場合は、ステップS6で変換の逆変換を施すことにより訓練データを増やしても、第1学習済みモデルM1の認識率が良くなる可能性がある。そこで、ステップS5で正しく分類されるようになった検証データ数と、ステップS5で正しく分類されなくなった検証データ数を比較して、ステップS5で正しく分類されるようになる検証データ数の差が大きい変換や、ステップS5で正しく分類されなくなる検証データがほとんど発生しない変換を優先して選択してもよい。

10

さらに、D1が正しく分類されるようになるほど大きな変化がない場合でも、正しいクラスである確率と正しくないクラスである確率が得られるので、それを比較するとD1の正しいクラスである確率が向上し、正しくないクラスである確率が低下する変換に着目し、逆に、D2、D3のように正しく分類されていた検証データについて、正しいクラスである確率が減少しないか減少が少ない変換を候補として選択してもよい。

【0037】

とくに、本実施形態では画像に対する変換を行うので、変換後のデータも画像として認識できる範囲のものであり、画像でないデータが生成されたり、人間が画像として認識できないようなデータが生成される可能性は低い。

20

【0038】

なお、このような効果は、いかなる学習データセットに対しても必ず成り立つというものではないが、少なくとも多くの学習データセットについては効率的に質の良いデータを追加することが可能である。

【0039】

実施の形態2 .

実施の形態2は、実施の形態1で説明した、基準データの選択方法および新たな訓練データの生成方法を変更するものである。以下、実施の形態1との相違点を説明する。

【0040】

図6に、実施の形態2における学習データの例を示す。検証データは検証データD4～D6を含む。図6の検証データD4～D6は[人物]のクラスに属するデータの例である。

30

【0041】

図7は、実施の形態2においてデータ生成装置10が実行する処理の流れを説明するフローチャートである。ステップS11およびS12は、図3のステップS1およびS2と同様である。

【0042】

ここでは、図6に示す検証データのうち、検証データD4およびD6は正しく分類されず(たとえば人物でないと認識され)、検証データD5は正しく分類された(たとえば人物であると認識された)とする。

40

【0043】

ステップS12の後、データ生成装置10は、第1学習済みモデルM1を用いて正しく分類された検証データのうちから、少なくとも1つの基準データを選択する(ステップS13)。実施の形態2では、実施の形態1と異なり、基準データは正しく分類された検証データのうちから選択される。ここでは、検証データD5が基準データとして選択されたものとする。

選択の基準は適宜設計可能であり、たとえば正しく分類された検証データのうちからランダムに選択するようにしてもよいし、正しく分類された検証データをすべて基準データとして選択するようにしてもよい。

50

【0044】

次に、データ生成装置10は、ステップS4と同様に、基準データに対して複数の異なる変換を施すことにより、複数の異なる変換済データを生成する(ステップS14)。

図8に、基準データに対する変換の例を示す。検証データD5に対してコントラストを緩和する変換を施すことにより、変換済データT4が生成される。同様に、検証データD5に対して縦方向に伸長する変換を施すことにより、変換済データT5が生成され、検証データD5に対して左右反転する変換を施すことにより、変換済データT6が生成される。

【0045】

次に、データ生成装置10は、ステップS5と同様に、第1学習済みモデルM1を用いて、変換済データT4~T6を分類するステップ(ステップS15)。ここでは、変換済データT5は正しく分類されず(たとえば人物でないと認識され)、変換済データT4およびT6は正しく分類された(たとえば人物であると認識された)ものとする。

10

【0046】

ここで、変換済データT4およびT6の分類を正しく行えた理由としては、訓練データ中に、変換済データT4のようにコントラストが弱い画像や、変換済データT6のように左右反転されたものに類似した画像が、訓練データ中に多く存在していた、という理由が考えられるとする。また、それにも関わらず、これらと同じ検証データD5から生成された変換済データT5の分類を誤ってしまった理由としては、訓練データ中に、変換済データT5のような縦方向に長い画像が不足していた、という理由が考えられるとする。このため、変換済データT5のように縦方向に長い画像を訓練データ中に追加して新たに機械学習をやり直せば、変換済データT5に似た画像を正しく分類できる学習済みモデルが生成される可能性が高まることが期待される。

20

【0047】

次に、データ生成装置10は、変換の内容に基づき、新たな訓練データを生成する(ステップS16)。より具体的には、変換済データのうち正しく分類されなかったものに施されていた変換を、少なくとも1つの訓練データに対して施すことにより、少なくとも1つの新たな訓練データを生成する。この例では、正しく分類されなかった変換済データT5に施されていた「縦方向に伸長する変換」という変換により、新たな訓練データを生成する。

30

【0048】

なお、新たな訓練データの元となる訓練データは、上述のように少なくとも1つであればよいが、すべての訓練データについて同様に新たな訓練データを生成してもよい。

【0049】

図9に、新たな訓練データの例を示す。訓練データK3に対して「縦方向に伸長する変換」を施すことにより、新たな訓練データN4が生成される。同様に、訓練データK4に対して「縦方向に伸長する変換」を施すことにより、新たな訓練データN5が生成され、訓練データK5に対して「縦方向に伸長する変換」を施すことにより、新たな訓練データN6が生成される。

【0050】

次に、データ生成装置10は、ステップS7と同様に、ステップS11で用いた訓練データと、ステップS16で生成された新たな訓練データとを用いて機械学習を行うことにより、第2学習済みモデルを生成する(ステップS17)。

40

【0051】

ここで、新たな訓練データN4~N6は、変換済データT5のように縦方向に長い画像となっているので、これらを用いて生成された第2学習済みモデルは、変換済データT4のようにコントラストが弱い画像や、変換済データT6のように左右反転したものに類似した画像のみならず、変換済データT5のように縦方向に長い画像についても、正しく分類できる可能性が高いといえる。すなわち、第2学習済みモデルは、より様々なデータを適切に分類できる頑強なモデルとなる。

50

【0052】

以上説明するように、本発明の実施の形態2に係るデータ生成装置10によれば、変換済データT5に類似する検証データ等を正しく分類するための質の良い訓練データを自動的に生成することができる。

【0053】

ステップS15では、検証データD5に異なる変換を施した変換済データT4～T5を分類することにより、ステップS16で施す変換の内容を決定している。正しく分類された検証データの全てまたは一部を選択し、ステップS15を実行する。ステップS16で施す変換は、ステップS15で、正しく分類されないデータが多い変換を、施す変換の内容と決定するように構成してもよい。このとき、一番正しく分類されないデータが多い変換だけを施してもよいし、上位複数の変換を施すように構成してもよい。

ステップS12で検証データD4、D6のように正しく分類されていなかった検証データが、ステップS15で正しく分類される場合はその検証データ数をカウントしてもよい。この検証データ数が一定値より多い場合は、ステップS16で変換を施すことにより訓練データを増やしても、第1学習済みモデルM1の認識率が良くなる可能性がある。そこで、ステップS12で正しく分類されなかった検証データ数と、ステップS15で正しく分類された検証データ数を比較して、ステップS15で正しく分類されなくなる検証データ数の差が大きい変換や、ステップS15で正しく分類されるようになる検証データがほとんど発生しない変換が有力と考えられる。

さらに、D5が正しく分類されなくなるほど大きな変化がない場合でも、正しいクラスである確率と正しくないクラスである確率が得られるので、それを比較するとD5の正しくないクラスである確率が向上し、正しいクラスである確率が低下する変換に着目し、逆に、D4、D6のように正しく分類されていなかった検証データについて、正しくないクラスである確率が減少しないか減少が少ない変換を候補として選択してもよい。

【0054】

実施の形態1で説明した処理を実行した後、実施の形態2で説明した処理を実行し、さらに実施の形態1で説明した処理に戻り実行する一連の流れを、基準データがなくなるまで繰り返してもよい。

実施の形態1および2において、以下のような変形を施すことができる。

ステップS1、S7、S11およびS17は、データ生成装置10以外のコンピュータが実行してもよい。また、ステップS7を実行せず、他の用途に新たな訓練データN2～N6を用いてもよい。

【0055】

データ変換の具体的内容は適宜設計可能であり、たとえば、任意の幾何学的線形変換を含むものであってもよい。幾何学的線形変換とは、たとえば直交2次元座標系において、座標に 2×2 行列を作用させる変換を意味する。幾何学的線形変換の例としては、回転、平行移動、拡大縮小、反転（たとえば左右反転）、これらの組み合わせ、等が挙げられる。拡大縮小は、1方向または2方向への拡大または縮小を含む。1方向への拡大または縮小は、縦方向の伸長、縦方向の圧縮、横方向の伸長、横方向の圧縮、斜め方向の伸長、斜め方向の圧縮、等を含む。2方向への拡大または縮小は、たとえば縦および横方向への伸長または圧縮（縦横で倍率が異なるものを含む）であるが、斜め方向への伸長または圧縮を含んでもよい。

【0056】

また、変換は、色値の変換を含むものであってもよい。色値の変換の例としては、明るさの変更、色調の変更、コントラストの強調または緩和、等が挙げられる。またニューラルネットで抽出した特徴量の追加や削除、ニューラルネットで抽出した特徴量の強調や弱める変換等であってもよい。

【0057】

学習データに含まれる訓練データと検証データは画像データ以外の種類のデータであってもよい。たとえば音声データであってもよい。この場合には、データの変換は、音の高

10

20

30

40

50

さの変更を含んでもよく、音声の再生スピードの変更を含んでもよい。また、データの変換は、雑音の付加または雑音の除去を含んでもよい。雑音の付加および除去ならびにそれぞれの逆変換は、当業者が任意に設計可能であるが、たとえば、2018年9月12日において <<https://tokyohappendix.com/audio-tips/audacity>> で公開されている技術を用いて行うことができる。実施の形態2においては、さらに、データの変換は、ローパスフィルタの適用（すなわち所定周波数以上の周波数成分を除去する）、ハイパスフィルタの適用（すなわち所定周波数以下の周波数成分を除去する）、等を含んでもよい。ローパスフィルタおよびハイパスフィルタの適用ならびにそれぞれの逆変換は、当業者が任意に設計可能であるが、たとえば、2018年9月12日において <https://pianoforte32.com/blog/files/DSEE_performance_of_HAP-Z1ES_by_pure_sound_source.html> で公開されている技術を用いて行うことができる。

10

またデータ変換は、ニューラルネットで抽出した特徴量の追加や削除、ニューラルネットで抽出した特徴量の強調や弱める変換等を含んでもよい。

【0058】

機械学習を行わせるためのモデルとしては、任意の形式のモデルを用いることができる。たとえばニューラルネットワークを用いてもよいし、他のモデルを用いてもよい。

【0059】

実施の形態1および2ではデータの分類を行うための学習済みモデル、訓練データおよび検証データについて説明したが、データの予測を行うための学習済みモデル、訓練データおよび検証データについても同様に実施が可能である。その場合には、学習済みモデルの出力値は予測値となる。また、出力値が適切であるか否かは、正しく分類されたか否かに代えて、予測精度が所定の閾値以上であったか否かに基づいて判定することができる。たとえば、予測された値と正解の値との間の二乗誤差を算出し、この二乗誤差が閾値未満であれば出力値が適切である（正しく予測された）と判定し、二乗誤差が閾値以上であれば出力値が不適切である（正しく予測されなかった）と判定してもよい。

20

【0060】

ステップS7およびS17において、訓練データ全体ではなく、訓練データの一部を拡張した場合、データ全体の傾向に影響を与えないように、新たな訓練データと、その元となった訓練データとについては、学習する際の重みの変更量を通常より小さくしてもよい。たとえば、通常の変更量（または係数）が w であり、1つの元の訓練データから n 個の新たな訓練データが生成された場合には、その元の訓練データと、その新たな訓練データとについては、重みの変更量（または係数）は $w/(n+1)$ としてもよい。具体例として、実施の形態1において、訓練データK1および訓練データK2と、これらから生成された新たな訓練データN2およびN3とについては、ステップS7においてこれらのデータを用いて学習する際の重みの変更量を通常の変更量の $1/2$ としてもよい。

30

【0061】

なお、上述した実施の形態1および2に係るデータ生成装置10における各機能は、処理回路によって実現される。各機能を実現する処理回路は、専用のハードウェアであってもよく、メモリに格納されるプログラムを実行するプロセッサであってもよい。図10は、本発明の実施の形態1および2に係るデータ生成装置10の各機能を専用のハードウェアである処理回路1000で実現する場合を示した構成図である。また、図11は、本発明の実施の形態1および2に係るデータ生成装置10の各機能をプロセッサ2001およびメモリ2002を備えた処理回路2000により実現する場合を示した構成図である。

40

【0062】

処理回路が専用のハードウェアである場合、処理回路1000は、例えば、単一回路、複合回路、プログラム化したプロセッサ、並列プログラム化したプロセッサ、ASIC (Application Specific Integrated Circuit)、FPGA (Field Programmable Gate Array)、またはこれらを組み合わせたものが該当する。データ生成装置10の各部の機能それぞれを個別の処理回路1000で実現してもよいし、各部の機能をまとめて処理回路1000で実現

50

してもよい。

【0063】

一方、処理回路がプロセッサ2001の場合、データ生成装置10の各部の機能は、ソフトウェア、ファームウェア、またはソフトウェアとファームウェアとの組み合わせにより実現される。ソフトウェアおよびファームウェアは、プログラムとして記述され、メモリ2002に格納される。プロセッサ2001は、メモリ2002に記憶されたプログラムを読み出して実行することにより、各部の機能を実現する。すなわち、データ生成装置10は、処理回路2000により実行されるときに、ステップS1~7またはS11~S17が結果的に実行されることになるプログラムを格納するためのメモリ2002を備える。

10

【0064】

これらのプログラムは、上述した各部の手順あるいは方法をコンピュータに実行させるものであるともいえる。ここで、メモリ2002とは、例えば、RAM(Random Access Memory)、ROM(Read Only Memory)、フラッシュメモリ、EPROM(Erasable Programmable Read Only Memory)、EEPROM(Electrically Erasable and Programmable Read Only Memory)等の、不揮発性または揮発性の半導体メモリが該当する。また、磁気ディスク、フレキシブルディスク、光ディスク、コンパクトディスク、ミニディスク、DVD等も、メモリ2002に該当する。

20

【0065】

なお、上述した各部の機能について、一部を専用のハードウェアで実現し、一部をソフトウェアまたはファームウェアで実現するようにしてもよい。

【0066】

このように、処理回路は、ハードウェア、ソフトウェア、ファームウェア、またはこれらの組み合わせによって、上述した各部の機能を実現することができる。

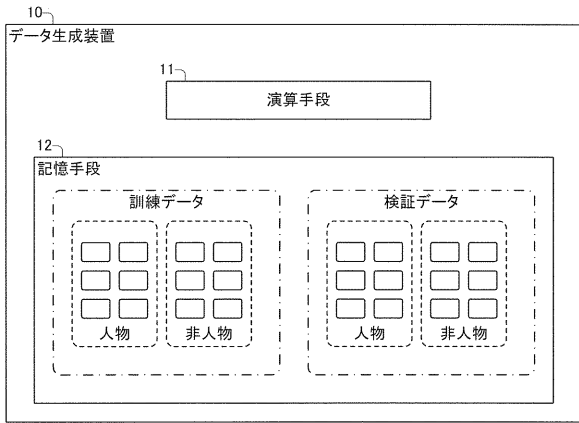
【符号の説明】

【0067】

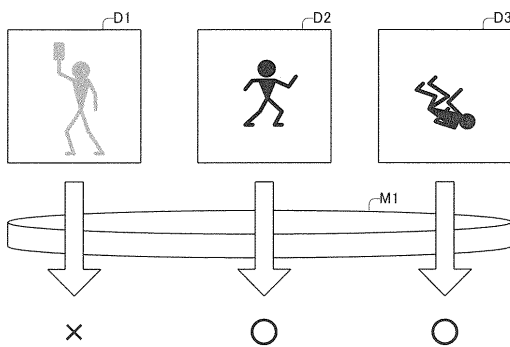
10 データ生成装置(コンピュータ)、D1~D6 検証データ(D1, D5 基準データ)、M1 第1学習済みモデル、K1~K5 訓練データ、N2~N6 新たな訓練データ、T1~T6 変換済データ。

30

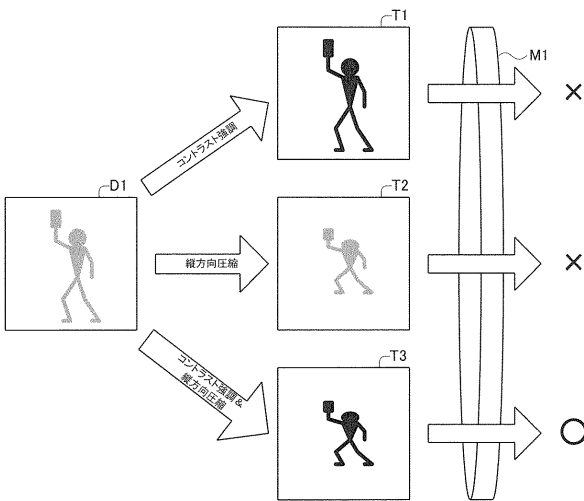
【 図 1 】



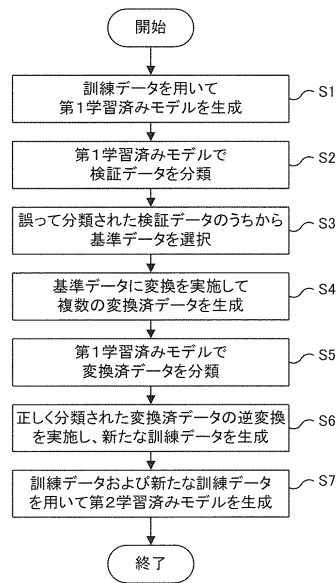
【 図 2 】



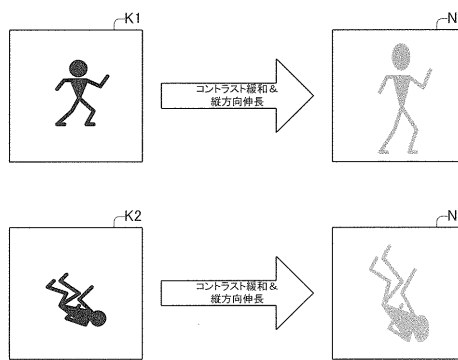
【 図 4 】



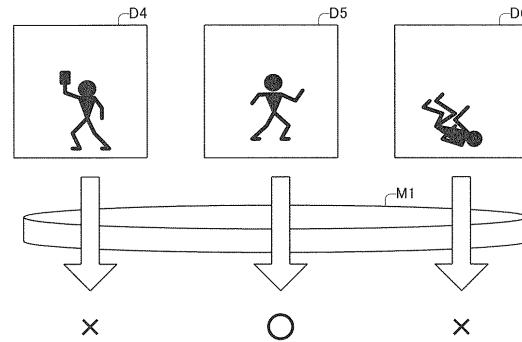
【 図 3 】



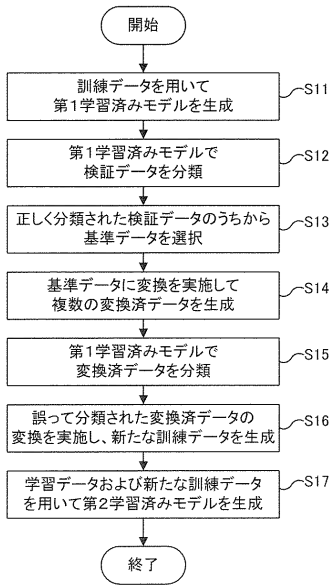
【 図 5 】



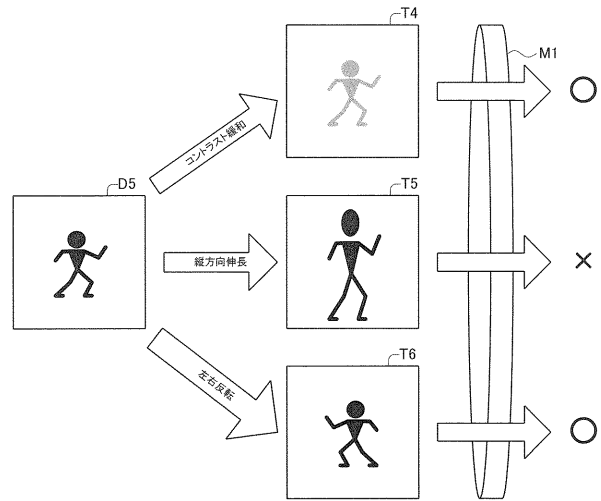
【 図 6 】



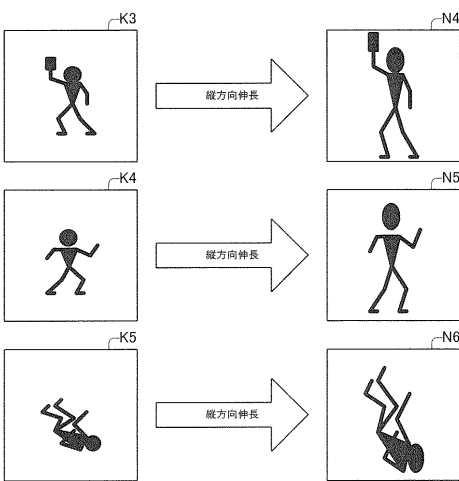
【 図 7 】



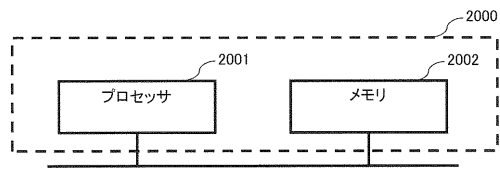
【 図 8 】



【 図 9 】



【 図 1 1 】



【 図 1 0 】

