【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】令和6年2月19日(2024.2.19)

【国際公開番号】WO2022/249392
【出願番号】特願2023-523860(P2023-523860)
【国際特許分類】
G06N 20/00(2019.01)
【FI】
G06N 20/00

【手続補正書】
【提出日】令和5年11月17日(2023.11.17)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
複数の訓練用例を取得する取得手段と、
用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練する訓練手段と、
前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する選択手段と、
前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する生成手段と、
を備えた情報処理装置。
【請求項２】
前記生成手段は、前記選択した訓練用例と、特徴量空間において前記選択した訓練用例の近傍に存在する用例とを合成して前記人工用例を生成する、請求項１に記載の情報処理装置。
【請求項３】
前記選択手段は、前記複数の訓練用例のうち、２つ以上の前記複数の予測結果にばらつきがある訓練用例を選択し、
前記生成手段は、２つ以上の前記選択した訓練用例を合成して前記人工用例を生成する、請求項１に記載の情報処理装置。
【請求項４】
前記生成手段は、
前記選択した訓練用例と、特徴量空間において前記選択した訓練用例の近傍に存在する用例とを合成する第１生成処理と、
２つ以上の前記選択した訓練用例を合成して前記人工用例を生成する第２生成処理と、
の何れかを実行することにより前記人工用例を生成する、請求項１に記載の情報処理装置。
【請求項５】
前記人工用例を前記複数の訓練用例に追加して、前記取得手段、前記訓練手段、前記選択手段、及び前記生成手段を再度機能させる、請求項１から４の何れか１項に記載の情報処理装置。
【請求項６】

前記生成手段は、
複数の前記人工用例を生成し、
複数の前記人工用例のうち類似条件を満たす２つの人工用例を１つの人工用例に統合する、請求項１から４の何れか１項に記載の情報処理装置。
【請求項７】
前記生成手段は、前記人工用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある人工用例を出力する、請求項１から６の何れか１項に記載の情報処理装置。
【請求項８】
前記機械学習モデル群は、前記人工用例を用いて訓練する訓練対象の機械学習モデルを含む、請求項１から７の何れか１項に記載の情報処理装置。
【請求項９】
複数の訓練用例を取得すること、
用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練すること、
前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択すること、及び、
前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成すること、
を含む情報処理方法。
【請求項１０】
コンピュータを情報処理装置として機能させるためのプログラムであって、前記コンピュータを、
複数の訓練用例を取得する取得手段と、
用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練する訓練手段と、
前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する選択手段と、
前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する生成手段と、
として機能させるプログラム。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、機械学習に用いる用例を生成する技術に関する。
【背景技術】
【０００２】
機械学習モデルの推論の精度は、その機械学習モデルを構築する際に用いた訓練用例の数や内容に依存することが知られている。機械学習モデルの推論精度を向上させるために、事前に用意された訓練用例から人工用例を生成することにより、訓練用例を増強する技術が知られている。例えば非特許文献１には、サポートベクタマシン（Support Vector Machine）の決定境界に最も近いマイノリティクラスのインスタンス（訓練用例）とその近傍のマイノリティクラスのインスタンスとを合成して、マイノリティクラスの仮想インスタンスを生成することが記載されている。
【先行技術文献】

【非特許文献】
【０００３】
【非特許文献１】Seyda Ertekin, “Adaptive Oversampling for Imbalanced Data Classification”, Information Sciences and Systems 2013", proceedings of the 28th International Symposium on Computer and Information Sciences (ISCIS), pp. 261－269), 2013
【発明の概要】
【発明が解決しようとする課題】
【０００４】
しかしながら、非特許文献１に記載の技術では、決定境界付近に仮想インスタンス（人工用例）が生成され、決定境界付近以外で訓練用例が不足している領域に人工用例が生成されない、という問題があった。
【０００５】
本発明の一態様は、上記の問題に鑑みてなされたものであり、その目的の一例は、機械学習の訓練に用いる訓練用例が不足している領域に人工用例を生成可能な技術を提供することである。
【課題を解決するための手段】
【０００６】
本発明の一側面に係る情報処理装置は、複数の訓練用例を取得する取得手段と、用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練する訓練手段と、前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する選択手段と、前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する生成手段と、を備える。
【０００７】
本発明の一側面に係る情報処理方法は、複数の訓練用例を取得すること、用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練すること、前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択すること、及び前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成すること、を含む。
【０００８】
本発明の一側面に係るプログラムは、コンピュータを情報処理装置として機能させるためのプログラムであって、前記コンピュータを、複数の訓練用例を取得する取得手段と、用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練する訓練手段と、前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する選択手段と、前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する生成手段と、として機能させる。
【発明の効果】
【０００９】
本発明の一態様によれば、機械学習の訓練に用いる訓練用例が不足している領域に人工用例を生成することができる。
【図面の簡単な説明】
【００１０】
【図１】本発明の例示的実施形態１に係る情報処理装置の構成を示すブロック図である。
【図２】本発明の例示的実施形態１に係る情報処理方法の流れを示すフロー図である。
【図３】本発明の例示的実施形態１に係る情報処理方法の具体例を模式的に示す図である。
【図４】本発明の例示的実施形態１により生成される人工用例を模式的に説明する図であ

る。
【図５】本発明の例示的実施形態２に係る情報処理装置の構成を示すブロック図である。
【図６】本発明の例示的実施形態２に係る情報処理方法の流れを示すフロー図である。
【図７】本発明の例示的実施形態２に係る情報処理方法の具体例を模式的に示す図である。
【図８】本発明の例示的実施形態２に係る第１生成処理の流れを示すフロー図である。
【図９】本発明の例示的実施形態２に係る第２生成処理の流れを示すフロー図である。
【図１０】本発明の例示的実施形態２に係る第３生成処理の流れを示すフロー図である。
【図１１】本発明の例示的実施形態３に係る情報処理方法の流れを示すフロー図である。
【図１２】本発明の例示的実施形態４に係る情報処理方法の流れを示すフロー図である。
【図１３】本発明の例示的実施形態５に係る情報処理方法を模式的に説明する図である。
【図１４】非特許文献１に記載の技術が生成する人工用例を模式的に説明する図である。
【図１５】本発明の例示的実施形態１～５に係る情報処理装置として機能するコンピュータの構成を示すブロック図である。
【発明を実施するための形態】
【００１１】
〔例示的実施形態１〕
本発明の第１の例示的実施形態について、図面を参照して詳細に説明する。本例示的実施形態は、後述する例示的実施形態の基本となる形態である。
【００１２】
＜情報処理装置の構成＞
本例示的実施形態に係る情報処理装置１０の構成について、図１を参照して説明する。図１は、情報処理装置１０の構成を示すブロック図である。情報処理装置１０は、複数の訓練用例から、機械学習モデル群を用いて人工用例を生成する装置である。
【００１３】
情報処理装置１０は、図１に示すように、取得部１１と、訓練部１２と、選択部１３と、生成部１４とを含む。取得部１１は、請求の範囲に記載した取得手段を実現する構成の一例である。訓練部１２は、請求の範囲に記載した訓練手段を実現する構成の一例である。選択部１３は、請求の範囲に記載した選択手段を実現する構成の一例である。生成部１４は、請求の範囲に記載した生成手段を実現する構成の一例である。
【００１４】
取得部１１は、複数の訓練用例を取得する。訓練部１２は、用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、複数の訓練用例を用いて訓練する。選択部１３は、複数の訓練用例のうち、訓練後の機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する。生成部１４は、複数の訓練用例のうち、選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する。
【００１５】
（機械学習モデル群）
機械学習モデル群は、複数の機械学習モデルを含む。各機械学習モデルは、用例を入力として予測結果を出力する。予測結果は、例えば、複数のラベルの各々が予測される予測確率を含むものであってもよい。この場合、最も予測確率が高いラベルを、予測結果と記載する場合もある。機械学習モデルは、一例として、決定木、ニューラルネットワーク、ランダムフォレスト、又はサポートベクタマシン等の機械学習アルゴリズムを用いて生成されたモデルである。ただし、各機械学習モデルの生成に用いられる機械学習アルゴリズムは、これらに限られない。複数の機械学習モデルは、全て同一の機械学習アルゴリズムを用いて生成されたモデルであってもよい。また、複数の機械学習モデルのうち少なくとも２つが、互いに異なる機械学習アルゴリズムを用いて生成されたモデルであってもよい。機械学習モデル群は、例えば情報処理装置１０のメモリに記憶されていてもよいし、情報処理装置１０と通信可能に接続された他の装置に記憶されていてもよい。
【００１６】

また、機械学習モデル群は、必ずしも全てが、「生成した人工用例を用いて訓練する訓練対象の機械学習モデル」でなくてもよい。換言すると、機械学習モデル群は、訓練対象である機械学習モデルの一部又は全部を含んでいてもよい。また、機械学習モデル群は、訓練対象である機械学習モデルを含んでいなくてもよい。訓練対象である機械学習モデルの数は複数であってもよく、また単数であってもよい。

【００１７】

（用例、訓練用例、人工用例）

用例は、各機械学習モデルに入力される情報であり、特徴量を含む。換言すると、用例は、特徴量空間に存在する。訓練用例は、機械学習モデル群の訓練に用いる用例である。訓練用例は、観測により得られる用例であってもよいし、人工的に生成された人工用例であってもよい。

【００１８】

（予測結果にばらつきがある訓練用例）

複数の予測結果にばらつきがある訓練用例とは、ばらつきの評価結果が「ばらつきが大きい」ことを示す訓練用例である。例えば、ばらつきの評価とは、複数の予測結果のばらつきが大きいか否かを評価することである。具体例として、ばらつきの評価は、投票結果のエントロピーに基づく評価であってもよい。投票結果のエントロピーについては、後述の例示的実施形態２で詳細を説明する。また、ばらつきの評価は、複数の予測結果のうち同一のラベルを示す予測結果の割合に基づく評価であってもよい。ただし、ばらつきの評価は、上述したものに限られない。以降、「複数の予測結果のばらつきが大きいと評価した訓練用例」を、「予測結果にばらつきがある訓練用例」とも記載する。また、「複数の予測結果のばらつきが大きくないと評価した訓練用例」を、「予測結果のばらつきが小さい訓練用例」とも記載する。

【００１９】

＜情報処理方法の流れ＞

本例示的実施形態に係る情報処理方法Ｓ１０の流れについて、図２を参照して説明する。図２は、情報処理方法Ｓ１０の流れを示すフロー図である。図２に示すように、情報処理方法Ｓ１０は、ステップＳ１０１～Ｓ１０４を含む。

【００２０】

（ステップＳ１０１）

ステップＳ１０１（取得処理）において、取得部１１は、複数の訓練用例を取得する。例えば、取得部１１は、複数の訓練用例をメモリから読み込むことにより取得してもよい。また、例えば、取得部１１は、複数の訓練用例を、入力装置から取得してもよいし、ネットワークを介して接続された装置から取得してもよい。本ステップで取得する複数の訓練用例は、観測用例及び人工用例の一方又は両方を含んでいる。

【００２１】

（ステップＳ１０２）

ステップＳ１０２（訓練処理）において、訓練部１２は、ステップＳ１０１で取得した複数の訓練用例を用いて、機械学習モデル群を訓練する。ここで、機械学習モデル群の各々の訓練に用いる訓練用例は、ステップＳ１０１で取得した複数の訓練用例の一部であってもよいし全部であってもよい。

【００２２】

（ステップＳ１０３）

ステップＳ１０３（選択処理）において、選択部１３は、複数の訓練用例のうち、予測結果にばらつきがある訓練用例を選択する。選択部１３は、そのような訓練用例を１つ選択してもよいし、複数選択してもよい。

【００２３】

具体的には、選択部１３は、複数の訓練用例のうち評価対象の訓練用例を、訓練後の各機械学習モデルに入力し、各機械学習モデルから出力される予測結果を取得する。これにより、選択部１３は、評価対象の訓練用例について複数の予測結果を得る。また、選択部

１３は、得られた複数の予測結果のばらつきを評価する。選択部１３は、複数の予測結果のばらつきが大きいと評価した場合、当該訓練用例を「予測結果にばらつきがある訓練用例」として選択する。

【００２４】

なお、選択部１３は、複数の訓練用例の一部又は全部を、ばらつきの評価対象とする。例えば、複数の訓練用例の一部を用いて機械学習モデル群の各々を訓練した場合、選択部１３は、複数の訓練用例の他の一部（すなわち、機械学習モデル群の訓練に用いなかった訓練用例）をそれぞれ評価対象としてもよい。

【００２５】

（ステップＳ１０４）

ステップＳ１０４（生成処理）において、生成部１４は、複数の訓練用例のうち、ステップＳ１０３で選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する。例えば、生成部１４は、選択した訓練用例と、特徴量空間においてその近傍に存在する他の訓練用例とを合成してもよい。また、例えば、生成部１４は、ステップＳ１０３において複数の訓練用例を選択した場合、選択した複数の訓練用例同士を合成してもよい。また、生成部１４は、２つの訓練用例を合成して１つの人工用例を生成してもよいし、３以上の訓練用例を合成して１つの人工用例を生成してもよい。また、生成部１４は、本ステップにおいて、１つの人工用例を生成してもよいし、複数の人工用例を生成してもよい。

【００２６】

（合成処理の具体例）

２つの訓練用例を合成して１つの人工用例を生成する場合、生成部１４が行う合成処理は、一例として、以下の（１）式で表される。

【数１】

$$\hat{x}_v = \lambda x_i + (1-\lambda)x_j \quad \cdots (1)$$

式（１）において、$\hat{x}_v$は人工用例を表し、x_iは、選択部１３が選択した訓練用例を表す。x_jは、選択部１３が選択した他の訓練用例であってもよいし、選択しなかった他の訓練用例であってもよい。選択しなかった他の訓練用例である場合、x_jは、x_iの近傍に存在する訓練用例である。λは、０≦λ≦１を満たす重み係数である。生成部１４は、一例として、係数λの値を、ランダム関数により発生させた乱数を用いて決定する。なお、生成部１４が行う合成処理は上述した手法に限られず、生成部１４は他の手法により複数の訓練用例を合成してもよい。

【００２７】

＜情報処理方法の具体例＞

情報処理方法Ｓ１０の具体例について、図３を参照して説明する。図３は、情報処理方法Ｓ１０の具体例を模式的に示す図である。

【００２８】

本具体例では、ステップＳ１０１において取得部１１が取得する複数の訓練用例Ｔは、訓練用例ｔ１，ｔ２，ｔ３，…を含む。ステップＳ１０２において訓練部１２が訓練する機械学習モデル群は、機械学習モデルｍ１，ｍ２，ｍ３，…を含む。機械学習モデルｍ１，ｍ２，ｍ３，…のそれぞれは、用例が入力されると、予測結果として「Ａ」及び「Ｂ」の何れかのラベルを出力する。機械学習モデルｍ１，ｍ２，ｍ３，…のそれぞれは、訓練用例Ｔの一部又は全部を用いて訓練される。ステップＳ１０３において、選択部１３は、評価対象の訓練用例ｔ１～ｔ１０について、複数の予測結果のばらつきを評価する。図３では、実線の円は、予測結果にばらつきがある訓練用例を示し、破線の円は、予測結果のばらつきが小さい訓練用例を示す。

【００２９】

具体的には、訓練用例ｔ１、ｔ２、ｔ５は、機械学習モデルｍ１，ｍ２，ｍ３，…から

得られる複数の予測結果が全て「Ａ」であり、この例では、予測結果のばらつきが大きくないと評価される。訓練用例ｔ６、ｔ９、ｔ１０は、機械学習モデルｍ１，ｍ２，ｍ３，…から得られる複数の予測結果が全て「Ｂ」であり、この例では、予測結果のばらつきが大きくないと評価される。訓練用例ｔ３、ｔ４は、機械学習モデルｍ１，ｍ２，ｍ３，…から得られる複数の予測結果のうち２つが「Ａ」であり、１つが「Ｂ」であり、この例では、予測結果のばらつきが大きいと評価される。訓練用例ｔ７、ｔ８は、機械学習モデルｍ１，ｍ２，ｍ３，…から得られる複数の予測結果のうち２つが「Ｂ」であり、１つが「Ａ」であり、この例では、予測結果のばらつきが大きいと評価される。

【００３０】

したがって、選択部１３は、予測結果にばらつきがある訓練用例ｔ３、ｔ４、ｔ７、ｔ８を選択する。ステップＳ１０４において、生成部１４は、予想結果にばらつきがある訓練用例ｔ３とその近傍の訓練用例ｔ５とを合成することにより、人工用例ｔ５１を生成する。また、生成部１４は、予想結果にばらつきがある訓練用例ｔ４とその近傍の訓練用例ｔ１とを合成することにより、人工用例ｔ５２を生成する。また、生成部１４は、複数の予想結果にばらつきがある訓練用例ｔ７、ｔ８同士を合成することにより、人工用例ｔ５３を生成する。図３では、二重線の円は、人工用例を示している。

【００３１】

＜本例示的実施形態の効果＞

本例示的実施形態は、複数の訓練用例を取得し、用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、複数の訓練用例を用いて訓練し、複数の訓練用例のうち、訓練後の機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択し、複数の訓練用例のうち、選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する、との構成を採用している。

【００３２】

ここで、複数の予想結果にばらつきがある訓練用例は、特徴量空間において訓練用例が不足している領域にあると考えられる。そのような訓練用例を含む複数の訓練用例を合成して得られる人工用例は、訓練用例が不足している領域に生成される可能性が高い。したがって、本例示的実施形態は、訓練用例が不足している領域に人工用例を生成可能である。

【００３３】

このような本例示的実施形態の効果について、図４、及び図１４を参照して説明する。図４は、本例示的実施形態が生成する人工用例を模式的に説明する図である。図１４は、非特許文献１に記載の技術が生成する人工用例を模式的に説明する図である。図４、及び図１４において、実線の円は、予測結果のばらつきが小さい訓練用例を示し、破線の円は、予測結果にばらつきがある訓練用例を示し、二重線の円は、人工用例を示す。領域Ｒ１、Ｒ２、Ｒ３は、特徴量空間における領域を示す。領域Ｒ１、Ｒ２、Ｒ３は、予測結果にばらつきがある人工用例を含んでおり、人工用例が不足している領域である。

【００３４】

図１４に示すように、非特許文献１に記載の技術は、サポートベクタマシンによる決定境界Ｂの近傍である領域Ｒ１に人工用例を生成する。しかしながら、非特許文献１に記載の技術は、決定境界Ｂの近傍以外で訓練用例が不足している領域Ｒ２、Ｒ３に人工用例を生成することが難しい。

【００３５】

これに対して、図４に示すように、本例示的実施形態は、予測結果にばらつきがある訓練用例を含む複数の訓練用例を合成して人工用例を生成する。したがって、本例示的実施形態は、訓練用例が不足している領域Ｒ１、Ｒ２、Ｒ３に、人工用例を生成することができる。また、本例示的実施形態は、一部の領域Ｒ１に偏って人工用例を生成することを抑制できる。

【００３６】

また、本例示的実施形態は、予測結果にばらつきがある訓練用例を選択するために、機

械学習モデル群を用いる、との構成を採用している。
【００３７】
これにより、本例示的実施形態は、非特許文献１に記載の技術のように決定境界の近傍に人工用例を生成する場合と比較して、偏った領域に人工用例を生成することを抑制することができる。
【００３８】
また、これにより、本例示的実施形態は、予測結果にばらつきがある訓練用例を選択するために予測確率を用いる場合と比較して、より訓練用例が不足している領域に人工用例を生成することができる。これは、例えば、機械学習モデル群に決定木が含まれる場合、決定木は、予測確率１で間違った予測をする可能性があるためである。
【００３９】
〔例示的実施形態２〕
本発明の第２の例示的実施形態について、図面を参照して詳細に説明する。なお、例示的実施形態１にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付し、その説明を適宜省略する。
【００４０】
＜情報処理装置の構成＞
本例示的実施形態に係る情報処理装置２０の構成について、図５を参照して説明する。図５は、情報処理装置２０の構成を示すブロック図である。情報処理装置２０は、複数の用例から、機械学習モデル群ＣＯＭ０を用いて人工用例を生成する装置である。
【００４１】
（機械学習モデル群）
機械学習モデル群ＣＯＭ０は、例示的実施形態１における機械学習モデル群とほぼ同様に構成される。ただし、本例示的実施形態では、機械学習モデル群ＣＯＭ０に含まれる複数の機械学習モデルは、そのうち少なくとも２つが、互いに異なる機械学習アルゴリズムを用いて生成されたものである。
【００４２】
また、機械学習モデル群ＣＯＭ０は、人工用例を用いて訓練する訓練対象である機械学習モデルを含む。また、機械学習モデル群ＣＯＭ０に含まれる機械学習モデルは、少なくとも１つが決定木である。ここでは、訓練対象である機械学習モデルが、決定木である。
【００４３】
情報処理装置２０は、図５に示すように、取得部２１と、訓練部２２と、選択部２３と、生成部２４と、ラベル付与部２５と、出力部２６と、制御部２７とを含む。取得部２１は、請求の範囲に記載した取得手段を実現する構成の一例である。訓練部２２は、請求の範囲に記載した訓練手段を実現する構成の一例である。選択部２３は、請求の範囲に記載した選択手段を実現する構成の一例である。生成部２４は、請求の範囲に記載した生成手段を実現する構成の一例である。ラベル付与部２５は、請求の範囲に記載したラベル付与手段を実現する構成の一例である。出力部２６は、請求の範囲に記載した出力手段を実現する構成の一例である。
【００４４】
取得部２１は、例示的実施形態１における取得部１１と同様に構成される。
【００４５】
訓練部２２は、例示的実施形態１における訓練部１２とほぼ同様に構成されるが、機械学習モデル群ＣＯＭ０を複数のグループに分けて訓練する点が異なる。訓練部２２による訓練処理の詳細については後述する。
【００４６】
選択部２３は、例示的実施形態１における選択部２３とほぼ同様に構成されるが、予測結果のばらつきの評価対象とする訓練用例の詳細が異なる。評価対象とする訓練用例の詳細については後述する。
【００４７】

生成部２４は、選択部２３が選択した訓練用例と、特徴量空間において選択した訓練用例の近傍に存在する用例とを合成して人工用例を生成する。生成部２４は、一例として、上記式（１）により２つの訓練用例を合成して人工用例を生成する。

【００４８】

ラベル付与部２５は、複数の訓練用例及び人工用例の一部又は全部にラベルを付与する。ラベル付与部２５は、一例として、ユーザ操作を受け付ける入力装置から出力される情報に基づきラベルを付与してもよい。また、一例として、ラベル付与部２５は、用例を入力としてラベルを出力するよう訓練された機械学習モデルに、訓練用例および人工用例を入力することにより得られるラベルを付与してもよい。この場合、ラベルを出力する機械学習モデルは、機械学習モデル群ＣＯＭ０に含まれる各機械学習モデルとは異なる機械学習モデルである。また、ラベルを出力する機械学習モデルは、機械学習モデル群に含まれる少なくとも１つの機械学習モデルより予測精度の高いモデルであることが望ましい。例えば、機械学習モデル群に含まれる、訓練対象の機械学習モデルが決定木であれば、ラベルを出力する機械学習モデルは、ランダムフォレストであってもよい。

【００４９】

出力部２６は、生成部２４が生成した人工用例を出力する。出力部２６は一例として、生成部２４が生成した人工用例を、外部記憶装置等の記録媒体に格納してもよい。また、出力部２６は一例として、表示装置等の出力装置に人工用例を出力してもよい。

【００５０】

制御部２７は、情報処理装置２０の各部を制御する。本例示的実施形態において制御部２７は特に、生成部２４が生成した人工用例を複数の訓練用例に追加して、取得部２１、訓練部２２、選択部２３、および生成部２４を再度機能させる。

【００５１】

＜情報処理方法の流れ＞

本例示的実施形態に係る情報処理方法Ｓ２０の流れについて、図６を参照して説明する。図６は、情報処理方法Ｓ２０の流れを示すフロー図である。

【００５２】

（ステップＳ２０１）

ステップＳ２０１（取得処理）において、取得部２１は、複数の訓練用例を取得する。取得する複数の訓練用例は、観測により得られた用例を含んでいてもよいし、人工用例を含んでいてもよい。

【００５３】

（ステップＳ２０２）

ステップＳ２０２において、ラベル付与部２５は、取得部２１が取得した複数の訓練用例の各々にラベルを付与する。

【００５４】

（ステップＳ２０３）

ステップＳ２０３（訓練処理）において、訓練部２２は、取得部２１が取得した複数の訓練用例の一部または全部を用いて、複数の機械学習モデル群の各々を訓練する。機械学習モデル群の各々を訓練する訓練処理の詳細については後述する。

【００５５】

（ステップＳ２０４）

ステップＳ２０４（選択処理）において、選択部２３は、取得部２１が取得した複数の訓練用例のうち、予測結果にばらつきがある訓練用例を１つ以上選択する。選択部２３が行う選択処理については後述する。予測結果にばらつきがある訓練用例を選択する処理の詳細については後述する。

【００５６】

（ステップＳ２０５）

ステップＳ２０５（生成処理）において、生成部２４は、選択部２３が選択した訓練用例を含む複数の訓練用例を、合成対象として特定する。また、生成部２４は、合成対象と

して特定した複数の訓練用例を合成して人工用例を生成する。生成部２４が行う生成処理の詳細については後述する。

【００５７】

（ステップＳ２０６）

ステップＳ２０６において、ラベル付与部２５は、生成部２４が生成した人工用例の各々にラベルを付与する。ステップＳ２０７において、制御部２７は、訓練処理を終了するかを判定する。制御部２７は、一例として、ステップＳ２０３～Ｓ２０６の処理を実行した回数が所定の閾値以上である場合、訓練処理を終了すると判定する。一方、ステップＳ２０３～Ｓ２０６の処理を実行した回数が所定の閾値未満である場合、訓練処理を終了しないと判定する。訓練処理を終了しない場合（ステップＳ２０７にてＮＯ）、制御部２７はステップＳ２０８の処理に進む。一方、訓練処理を終了する場合（ステップＳ２０７にてＹＥＳ）、制御部２７はステップＳ２０９の処理に進む。

【００５８】

（ステップＳ２０８）

ステップＳ２０８において、制御部２７は、これまでに実行したステップＳ２０６で生成された１以上の人工用例を複数の訓練用例に追加する。ステップＳ２０８の処理を終えると、制御部２７は、ステップＳ２０３の処理に戻る。換言すると、制御部２７は、人工用例を複数の訓練用例に追加して、取得部２１、訓練部２２、選択部２３、および生成部２４を再度機能させる。

【００５９】

（ステップＳ２０９）

ステップＳ２０９において、出力部２６は、これまでに実行したステップＳ２０６で生成された１以上の人工用例を出力する。

【００６０】

＜訓練対象の機械学習モデルの訓練＞

このようにして情報処理方法Ｓ２０を用いて生成された１つ以上の人工用例は、訓練対象の機械学習モデルを訓練するために用いられる。訓練対象の機械学習モデルを訓練する処理は、例えば、訓練部２２が実行してもよい。

【００６１】

（訓練処理、選択処理の具体例）

ステップＳ２０３～Ｓ２０４における訓練処理及び選択処理の具体例について、図７を参照して説明する。図７は、情報処理方法Ｓ２０の具体例を模式的に示す図である。

【００６２】

図７に示すように、ステップＳ２０３において、訓練部２２は、機械学習モデル群ＣＯＭ０を、複数のグループＣＯＭｉ（ｉ＝１，２，…，Ｍ、Ｍは２以上の整数）に分割して訓練を行う。以降、分割した各グループを、機械学習モデル群ＣＯＭｉと記載する。また、機械学習モデル群ＣＯＭｉには、複数の機械学習モデルｍｉ－ｊ（ｊ＝１，２，…）が含まれる。以降、機械学習モデル群ＣＯＭｉに含まれる複数の機械学習モデルを、機械学習モデルｍｉ－ｊと記載する。機械学習モデル群ＣＯＭｉに含まれる複数の機械学習モデルｍｉ－ｊは、全てが同一の機械学習アルゴリズムにより生成されたモデルであってもよいし、そのうち少なくとも２つが互いに異なる機械学習アルゴリズムにより生成されたモデルであってもよい。また、機械学習モデル群ＣＯＭｉ１に含まれる機械学習モデルｍｉ１－ｊの個数は、機械学習モデル群ＣＯＭｉ２に含まれる機械学習モデルｍｉ２－ｊの個数と同一であってもよいし、異なっていてもよい（ｉ１＝１，２，…，Ｍ、ｉ２＝１，２，…，Ｍ、ｉ１≠ｉ２）。

【００６３】

ステップＳ２０３において、訓練部２２は、ステップＳ２０１において取得部２１が取得した訓練用例群Ｔから、訓練用例群Ｄｉを抽出する。訓練用例群Ｄｉは、訓練用例群Ｔの一部である。例えば、訓練部２２は、ランダムサンプリングにより訓練用例群Ｄｉを抽出してもよい。訓練用例群Ｄｉ１とＤｉ２とは、含まれる訓練用例が全て同一であっても

よいし、一部または全部が異なっていてもよい。訓練部２２は、訓練用例群Ｄｉを用いて、機械学習モデル群ＣＯＭｉに含まれる各機械学習モデルｍｉ－ｊを訓練することを、ｉ＝１，２，…，Ｍについて繰り返す。

【００６４】

ステップＳ２０４において、選択部２３は、機械学習モデル群ＣＯＭｉを用いて、予測結果にばらつきがある訓練用例を選択することを、ｉ＝１，２，…，Ｍについて繰り返す。。具体的には、選択部２３は、機械学習モデル群ＣＯＭｉの訓練に用いなかった各訓練用例（すなわち、訓練用例群Ｔのうち訓練用例群Ｄｉ以外）について、予測結果のばらつきを評価する。これにより、選択部２３は、このような評価対象の訓練用例のうち、予測結果予測結果にばらつきがある訓練用例を選択する。図７の例では、選択部２３は、機械学習モデル群ＣＯＭ１を用いて、予測結果にばらつきがある訓練用例ｔ１，ｔ２，…を選択している。また、選択部２３は、機械学習モデル群ＣＯＭ２を用いて、予測結果にばらつきがある訓練用例ｔ１１，ｔ１２，…を選択している。

【００６５】

選択部２３は例えば、評価対象の各訓練用例について、ＱＢＣ（query by committee）の手法における投票結果のエントロピー（ｖｏｔｅ　ｅｎｔｒｏｐｙ）の指標を用いて、ばらつきの評価を行う。例えば、以下の式（２）は、投票結果のエントロピーが最大である訓練用例 $\hat{}$ ｘを示す式である。

【数２】

$$\hat{x} = \underset{x}{\operatorname{argmax}} \left(-\sum_{y} \frac{V(y)}{C} \log \frac{V(y)}{C} \right) \quad \cdots (2)$$

式（２）において、Ｃは、機械学習モデル群ＣＯＭｉにおける機械学習モデルｍｉ－ｊの総数を示す。Ｖ（ｙ）は、機械学習モデル群ＣＯＭｉにおいてラベルｙを予測した機械学習モデルｍｉ－ｊの数を示す。選択部２３は、式（２）が示す訓練用例 $\hat{}$ ｘを、予測結果にばらつきがある訓練用例として選択してもよい。この場合、各機械学習モデル群ＣＯＭｉについて選択部２３が選択する、予測結果にばらつきがある訓練用例の数は１つである。換言すると、この場合、選択部２３は、Ｍ個の機械学習モデル群ＣＯＭｉを用いて、予測結果にばらつきがある訓練用例をＭ個選択する。また、選択部２３は、各機械学習モデル群ＣＯＭｉについて、投票結果のエントロピーが大きい順に所定数の訓練用例を選択してもよいし、投票結果のエントロピーが閾値以上の訓練用例を選択してもよい。この場合、各機械学習モデル群ＣＯＭｉについて選択部２３が選択する、予測結果にばらつきがある訓練用例の数は複数でありうる。換言すると、この場合、選択部２３は、Ｍ個の機械学習モデル群ＣＯＭｉを用いて、予測結果にばらつきがある訓練用例をＭ個以上選択する。さらに、選択部２３は、このようにして選択した、予測結果にばらつきがあるＭ個以上の訓練用例の中から、ランダムに１つ又は所定数を選択してもよいし、投票結果のエントロピーが大きい順に１つ又は所定数を選択してもよい。

【００６６】

（生成処理の具体例）

ステップＳ２０５における生成処理の具体例について説明する。ステップＳ２０５において、生成部２４は、予測結果にばらつきがある訓練用例を用いて、第１生成処理Ｓ３０、第２生成処理Ｓ４０、及び第３生成処理Ｓ５０の何れかを実行することにより、人工用例を生成する。第１生成処理Ｓ３０は、予測結果にばらつきがある訓練用例とその近傍の訓練用例とを合成して人工用例を生成する処理である。第２生成処理は、予測結果にばらつきがある２つ以上の訓練用例を合成して人工用例を生成する処理である。第３生成処理は、第１生成処理及び第２生成処理の何れかを選択的に実行する処理である。

【００６７】

（第１生成処理）

第１生成処理Ｓ３０について、図８を参照して説明する。図８は、第１生成処理Ｓ３０の流れを示すフロー図である。図８において、第１生成処理Ｓ３０は、ステップＳ３０１～Ｓ３０２を含む。ここで、先行して実施されたステップＳ２０４では、予測結果にばらつきがある１または複数の訓練用例が選択されている。生成部２４は、予測結果にばらつきがある１または複数の訓練用例のそれぞれ（以下では、当該訓練用例と記載）について、以下のステップＳ３０１～Ｓ３０２を実行する。

【００６８】

（ステップＳ３０１）

ステップＳ３０１において、生成部２４は、当該訓練用例の近傍の訓練用例を選択する。近傍の訓練用例は、予測結果にばらつきがある訓練用例であってもよいし、予測結果のばらつきが小さい訓練用例であってもよい。例えば、近傍の訓練用例は、訓練用例群Ｔのうち、当該訓練用例との特徴量空間における距離が最も近い訓練用例であってもよい。また、例えば、近傍の訓練用例は、訓練用例群Ｔのうち、当該訓練用例との特徴量空間における距離が閾値以下の訓練用例であってもよい。

【００６９】

（ステップＳ３０２）

ステップＳ３０２において、生成部２４は、当該訓練用例と、ステップＳ３０１で選択した近傍の訓練用例とを合成して人工用例を生成する。例えば、図７の例では、予測結果にばらつきがある訓練用例ｔ１とその近傍の訓練用例とを合成して人工用例ｔｖ１－１が生成される。また、予測結果にばらつきがある訓練用例ｔ２とその近傍の訓練用例とを合成して人工用例ｔｖ１－２が生成される。

【００７０】

ここで、生成部２４は、合成処理の一例として、上記式（１）を用いてもよい。また、生成部２４は、合成処理の他の例として、ＭＵＮＧＥ（参考文献１参照）、ＳＭＯＴＥ（参考文献２参照）等の公知の技術を用いてもよい。

【００７１】

［参考文献１］ Bucilua, C., Caruana, R. and Niculescu-Mizil, A., “Model Compression”, Proc. ACM SIGKDD, pp. 535-541 (2006)

［参考文献２］ Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., “SMOTE: Synthetic minority over-sampling technique”, Journal of Artificial Intelligent Research, 16, 321-357 (2002).

（第２生成処理）

第２生成処理Ｓ４０について、図９を参照して説明する。図９は、第２生成処理Ｓ４０の流れを説明するフロー図である。図９に示すように、第２生成処理Ｓ４０は、ステップＳ４０１～Ｓ４０２を含む。なお、第２生成処理は、ステップＳ２０４において、予測結果にばらつきがある訓練用例が複数選択されている場合に実行可能である。生成部２４は、生成部２４は、予測結果にばらつきがある複数の訓練用例のそれぞれ（以下、当該訓練用例と記載）について、以下のステップＳ４０１～Ｓ４０２を実行する。

【００７２】

（ステップＳ４０１）

ステップＳ４０１において、選択部２３は、予測結果にばらつきがある複数の訓練用例のうち、当該訓練用例とは異なる他の訓練用例を選択する。換言すると、選択部２３は、複数の訓練用例のうち、複数の予測結果にばらつきがある２つ以上の訓練用例を選択する。例えば、選択部２３は、そのような他の訓練用例を、予測結果にばらつきがある複数の訓練用例からランダムに選択してもよい。また、例えば、選択部２３は、そのような他の訓練用例として、予測結果にばらつきがある複数の訓練用例のうち当該訓練用例との特徴量空間における距離が最も小さいもの、又は距離が閾値以下のものを選択してもよい。なお、当該訓練用例が、既に合成に用いられている場合、当該訓練用例に関するステップＳ４０１～Ｓ４０２の処理は、実行されなくてもよい。

【００７３】

（ステップＳ４０２）

ステップＳ４０２において、生成部２４は、当該訓練用例と、ステップＳ４０１で選択した他の訓練用例とを合成して人工用例を生成する。例えば、図７の例では、予測結果にばらつきがある訓練用例ｔ１１及びｔ１２を合成して人工用例ｔｖ２－１が生成される。ここで、生成部２４が合成する２以上の訓練用例は、この例のように、同一の機械学習モデル群ＣＯＭｉを用いて選択したものであってもよいし、そのうち少なくとも１つが他とは異なる機械学習モデル群ＣＯＭｉを用いて選択したものであってもよい。例えば、図７の例では、生成部２４は、予測結果にばらつきのある訓練用例ｔ１、ｔ２、…、ｔ１１、ｔ１２、…の中から２つ以上の訓練用例を選択し、選択した訓練用例を合成して人工用例ｔｖ１－１、ｔｖ１－２、ｔｖ２－１、又はｔｖ２－２を生成してもよい。なお、ステップＳ４０２における合成処理に用いる手法については、ステップＳ３０２で説明した通りであるため、詳細な説明を繰り返さない。

【００７４】

（第３生成処理）

第３生成処理について、図１０を参照して説明する。図１０は、第３生成処理Ｓ５０の流れを説明するフロー図である。図１０において、第３生成処理Ｓ５０は、ステップＳ５０１～Ｓ５０３を含む。ここで、ステップＳ２０４では、予測結果にばらつきがある１または複数の訓練用例が選択されている。生成部２４は、予測結果にばらつきがある１または複数の訓練用例のそれぞれ（以下では、当該訓練用例と記載）について、以下のステップＳ５０１～Ｓ５０３を実行する。

【００７５】

（ステップＳ５０１）

ステップＳ５０１において、生成部２４は、第１生成処理及び第２生成処理の何れかを選択する。例えば、生成部２４は、ランダム関数により決定した確率ｐを用いて第１生成処理を選択し、第１生成処理を選択しなかった場合に第２生成処理を選択してもよい。なお、第１生成処理及び第２生成処理の何れかを選択する手法は、確率ｐを用いる手法に限らず、他の手法であってもよい。

【００７６】

（ステップＳ５０２～Ｓ５０４）

ステップＳ５０２において、生成部２４は、いずれを選択したかを判定する。第１生成処理を選択した場合、生成部２４はステップＳ５０３の処理に進み、第１生成処理を実行する。一方、第２生成処理を選択した場合、生成部２４はステップＳ５０４の処理に進み、第２生成処理を実行する。第１生成処理及び第２生成処理の詳細については、上述した通りである。

【００７７】

＜本例示的実施形態の効果＞

本例示的実施形態は、予測結果にばらつきがある訓練用例と、その近傍にある訓令用例とを合成して人工用例を生成する第１生成処理を実行する、との構成を有する。

【００７８】

ここで、第１生成処理により生成される人工用例は、予測結果にばらつきがある訓練用例の近傍に生成される。予測結果にばらつきがある訓練用例は、特徴量空間において訓練用例が不足している領域にあると考えられる。したがって、このような人工用例は、訓練用例が不足している領域に生成される。

【００７９】

また、本例示的実施形態は、予測結果にばらつきがある２つ以上の訓練用例を合成して人工用例を生成する第２生成処理を実行する、との構成を有する。

【００８０】

ここで、第２生成処理により生成される人工用例は、訓練用例が不足している領域にある訓練用例同士が合成されたものである。したがって、このような人工用例が存在する領域も、訓練用例が不足している可能性が高い。

【００８１】
また、本例示的実施形態は、第１生成処理と、第２生成処理との何れかを選択して実行することにより人工用例を生成する第３生成処理を実行する、との構成を有する。
【００８２】
ここで、例えば、第３生成処理により生成される人工用例は、第１生成処理又は第２生成処理により生成される。第１生成処理により人工用例が生成される領域と、第２生成処理により人工用例が生成される領域とは、異なり得る。したがって、第３生成処理により複数の人工用例を生成する場合、これらの人工用例は、訓練用例が不足しているより多様な領域に分散して生成される可能性が高くなる。
【００８３】
その結果、本例示的実施形態は、第１生成処理、第２生成処理、及び第３生成処理の何れかを実行することにより、訓練用例が十分である領域に偏って過剰に人工用例を生成することがなく、訓練用例がより不足している領域に人工用例を生成することができる。
【００８４】
また、本例示的実施形態は、機械学習モデル群が、訓練対象の機械学習モデルを含んでいる。これにより、本例示的実施形態は、訓練対象の機械学習モデルの精度向上により効果的な人工用例を生成することができる。
【００８５】
また、本例示的実施形態は、機械学習モデル群のうち少なくとも２つが、互いに異なる機械学習アルゴリズムにより生成されるモデルである、との構成を採用している。
【００８６】
これにより、予測結果にばらつきがある訓練用例をより精度よく選択することができる。
【００８７】
また、本例示的実施形態は、訓練対象の機械学習モデルが決定木であり、サポートベクタマシンではない。また、本例示的実施形態は、そのような訓練対象の機械学習モデルが機械学習モデル群ＣＯＭ０に含まれている。したがって、サポートベクタマシンの決定境界近傍に人工用例を生成する非特許文献１に記載の技術と比較して、本例示的実施形態は、訓練対象の機械学習モデルの精度向上により効果的な人工用例を生成することができる。
【００８８】
〔例示的実施形態３〕
本発明の例示的実施形態３について、図面を参照して詳細に説明する。なお、例示的実施形態２にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付記してその説明を繰り返さない。本例示的実施形態は、例示的実施形態２における生成部２４を次のように変形した形態である。
【００８９】
＜生成部の構成＞
本例示的実施形態において、生成部２４は、複数の人工用例を生成する。また、生成部２４は、生成した複数の人工用例のうち、類似条件を満たす２つの人工用例を１つの人工用例に統合する。ここで、類似条件とは、用例が類似することを示す条件である。類似条件は、例えばコサイン類似度が閾値以上であることであってもよいし、特徴量空間における距離が閾値以下であることであってもよい。ただし、類似条件はこれらに限られない。統合する処理の詳細については後述する。
【００９０】
＜情報処理方法の流れ＞
本例示的実施形態における情報処理方法Ｓ２０Ａについて、図１１を参照して説明する。図１１は、例示的実施形態３に係る情報処理方法Ｓ２０Ａの流れを説明するフロー図である。図１１に示す情報処理方法Ｓ２０Ａは、例示的実施形態２に係る情報処理方法Ｓ２０とほぼ同様に構成されるが、ステップＳ２０５Ａをさらに含む点が異なる。

【００９１】
（ステップＳ２０５Ａ）
ステップＳ２０５Ａにおいて、生成部２４は、ステップＳ２０５において生成した人工用例のうち、類似する２つの人工用例を統合する。具体的には、生成部２４は、今回のステップＳ２０５において生成した人工用例と、前回までのステップＳ２０５において生成した人工用例の何れかとが類似条件を満たすか否かを判定する。類似条件を満たすと判定した場合、生成部２４は、類似条件を満たす２つの人工用例を統合する。
【００９２】
（統合処理の具体例）
統合処理の一例として、２つの人工用例を合成する処理が挙げられる。この場合、生成部２４は、２つの人工用例を合成して１つの人工用例を生成し、類似条件を満たした元の２つの人工用例を削除する。また、統合処理の他の例として、２つの人工用例のうち一方を削除する処理が挙げられる。なお、統合処理は、類似条件を満たす２つの人工用例の代わりに、当該２つの人工用例を参照して生成した１つの人工用例を採用する処理であればよく、上述した処理に限られない。なお、人工用例を削除するとは、ステップＳ２０６でラベルを付与する対象、及びステップＳ２０８で訓練用例に追加する対象から削除することである。これにより、統合された人工用例に対して、ラベルが付与されるとともに訓練用例に追加される。
【００９３】
＜本例示的実施形態の効果＞
本例示的実施形態においては、生成部が、複数の人工用例を生成し、生成した複数の人工用例のうち、類似条件を満たす２つの人工用例を１つの人工用例に統合する、との構成が採用されている。
【００９４】
ここで、訓練用例が不足している領域に存在する複数の用例が類似する場合、それらの用例を用いて機械学習モデルを訓練することは、機械学習モデルの精度向上において効率的ではない。したがって、本例示的実施形態は、類似条件を満たす人工用例を統合することにより、訓練用例が不足している領域に、より効率的に機械学習モデルの精度を向上させることができる人工用例を生成することができる。
【００９５】
〔例示的実施形態４〕
本発明の例示的実施形態４について、図面を参照して詳細に説明する。なお、例示的実施形態２にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付記してその説明を繰り返さない。本例示的実施形態は、例示的実施形態２における生成部２４を次のように変形した形態である。
【００９６】
＜生成部の構成＞
本例示的実施形態において、生成部２４は、生成した人工用例のうち、訓練後の機械学習モデル群ＣＯＭ０を用いて得られる複数の予測結果にばらつきがある人工用例を出力する。ここで、ばらつきがある人工用例は、ばらつきの評価結果が「ばらつきが大きい」ことを示す人工用例である。ばらつきの評価の詳細については、上述した通りであるため、詳細を繰り返さない。換言すると、生成部２４は、生成した人工用例のばらつきを、訓練後の機械学習モデル群ＣＯＭ０を用いて事後評価し、事後評価により予測結果にばらつきがある人工用例を採用する。
【００９７】
＜情報処理方法の流れ＞
本例示的実施形態における情報処理方法Ｓ２０Ｂについて、図１２を参照して説明する。図１２は、例示的実施形態４に係る情報処理方法Ｓ２０Ｂの流れを説明するフロー図である。図１２に示す情報処理方法Ｓ２０Ｂは、例示的実施形態２に係る情報処理方法Ｓ２０とほぼ同様に構成されるが、ステップＳ２０５Ｂをさらに含む点が異なる。

【００９８】
（ステップＳ２０５Ｂ）
ステップＳ２０５Ｂにおいて、生成部２４は、ステップＳ２０５において生成した人工用例を事後評価する。
【００９９】
具体的には、生成部２４は、当該人工用例について、機械学習モデル群ＣＯＭ０を用いて予測結果のばらつきを評価する。例えば、図７に示した例では、生成部２４は、人工用例ｔｖ１－１について、機械学習モデル群ＣＯＭ１を用いて予測結果のばらつきを評価する。このように、ばらつきの評価に用いる機械学習モデル群ＣＯＭ１は、当該人工用例ｔｖ１－１を生成するために参照した訓練用例ｔ１の評価に用いたものであることが望ましい。機械学習モデル群ＣＯＭ０を用いて予測結果のばらつきを評価する処理の詳細については、例示的実施形態２において説明した通りである。
【０１００】
生成部２４は、ステップＳ２０５において生成した人工用例について、予測結果のばらつきが大きくないと評価した場合には、当該人工用例を削除する。ここで、人工用例を削除するとは、ステップＳ２０６でラベルを付与する対象、及びステップＳ２０８で訓練用例に追加する対象から削除することである。これにより、予測結果にばらつきがある人工用例に対してラベルが付与されるとともに、訓練用例に追加される。
【０１０１】
＜本例示的実施形態の効果＞
本例示的実施形態においては、生成部が、生成した人工用例のうち、訓練後の機械学習モデル群を用いて得られる複数の予測結果にばらつきがある人工用例を出力する、との構成が採用されている。
【０１０２】
ここで、予測結果にばらつきがある訓練用例を含む複数の訓練用例を合成して得られた人工用例は、必ずしも予測結果にばらつきがあるとは限らない。換言すると、このようにして生成した人工用例は、予測結果のばらつきが小さい可能性がある。予測結果のばらつきが小さい訓練用例を用いて機械学習モデルを訓練することは、機械学習モデルの精度向上において効率的ではない。したがって、本例示的実施形態は、生成した人工用例を事後評価することにより、訓練用例が不足している領域に、より効率的に機械学習モデルの精度を向上させることができる人工用例を生成することができる。
【０１０３】
〔例示的実施形態５〕
本発明の例示的実施形態５について、図面を参照して詳細に説明する。なお、例示的実施形態２にて説明した構成要素と同じ機能を有する構成要素については、同じ符号を付記してその説明を繰り返さない。
【０１０４】
本例示的実施形態は、例示的実施形態２における機械学習モデル群ＣＯＭ０の構成、及び情報処理方法Ｓ２０におけるステップＳ２０３～Ｓ２０４を次のように変形した形態である。本例示的実施形態について、図１３を参照して説明する。図１３は、本例示的実施形態に係る情報処理方法を模式的に説明する図である。
【０１０５】
（機械学習モデル群）
図１３に示すように、本例示的実施形態では、機械学習モデル群ＣＯＭ０は、機械学習モデルｍｊ（ｊ＝１，２，…，Ｍ）を含む。各機械学習モデルｍｊは、同一の機械学習アルゴリズムによって生成されるモデルである。例えば、各機械学習モデルｍｊは、決定木であってもよい。
【０１０６】
（ステップＳ２０３）
本例示的実施形態のステップＳ２０３において、訓練部２２は、ステップＳ２０１で取

得部２１が取得した訓練用例群Ｔから、訓練用例群Ｄｊを抽出する。訓練用例群Ｄｊは、訓練用例群Ｔの一部である。例えば、訓練部２２は、ランダムサンプリングにより訓練用例群Ｄｊを抽出してもよい。訓練部２２は、訓練用例群Ｄｊを用いて、機械学習モデルｍｊを訓練することを、ｊ＝１，２，…，Ｍについて繰り返す。

【０１０７】

ここで、訓練用例群Ｄｊ１とＤｊ２とは、含まれる訓練用例が全て同一であってもよいが、一部または全部が異なることが望ましい（ｊ１＝１，２，…，Ｍ、ｊ２＝１，２，…，Ｍ、ｊ１≠ｊ２）。少なくとも一部が異なる訓練用例群Ｄｊ１とＤｊ２とをそれぞれ用いることにより、機械学習モデル群ｍｊ１及びｍｊ２は、それぞれを構成するパラメータが互いに異なるように訓練される。

【０１０８】

（ステップＳ２０４）

本例示的実施形態のステップＳ２０４において、選択部２３は、機械学習モデル群ＣＯＭ０を用いて、訓練用例群Ｔに含まれる各訓練用例について、予測結果のばらつきを評価する。また、選択部２３は、予測結果にばらつきがある訓練用例を選択する。図１３の例では、選択部２３は、機械学習モデル群ＣＯＭ０を用いて、予測結果にばらつきがある訓練用例ｔ１，ｔ３，…を選択している。

【０１０９】

ステップＳ２０５の処理は、例示的実施形態２で説明した通りである。すなわち、図１３の例では、予測結果にばらつきがある訓練用例ｔ１、ｔ２、…の各々について、第１生成処理、第２生成処理、及び第３生成処理の何れかが実行される。これにより、人工用例ｔｖ１、ｔｖ２、…が生成される。

【０１１０】

＜本例示的実施形態の効果＞

本例示的実施形態は、機械学習モデル群を構成する機械学習モデルとして、全て同一の機械学習アルゴリズムによって生成されたモデルを用い、取得した訓練用例群の中から予測結果にばらつきがある訓練用例を選択する、との構成を採用している。

【０１１１】

これにより、本例示的実施形態は、取得した訓練用例群の全てに亘って訓練用例が不足している領域に、人工用例を生成することができる。

【０１１２】

また、本例示定期実施形態は、機械学習モデル群に含まれる機械学習モデルが全て決定木である場合、当該機械学習モデルの精度向上により効果的な人工用例を生成することができる。その理由について説明する。決定木は、訓練用例の小さな変更に対して木の構造が大きく変化しうる。そのため、複数の決定木を含む機械学習モデル群を用いることにより、予測結果にばらつきがある訓練用例をより精度よく選択することができるためである。

【０１１３】

〔ソフトウェアによる実現例〕

情報処理装置１０，２０の一部又は全部の機能は、集積回路（ＩＣチップ）等のハードウェアによって実現してもよいし、ソフトウェアによって実現してもよい。

【０１１４】

後者の場合、情報処理装置１０，２０は、例えば、各機能を実現するソフトウェアであるプログラムの命令を実行するコンピュータによって実現される。このようなコンピュータの一例（以下、コンピュータＣと記載する）を図１５に示す。コンピュータＣは、少なくとも１つのプロセッサＣ１と、少なくとも１つのメモリＣ２と、を備えている。メモリＣ２には、コンピュータＣを情報処理装置１０，２０として動作させるためのプログラムＰが記録されている。コンピュータＣにおいて、プロセッサＣ１は、プログラムＰをメモリＣ２から読み取って実行することにより、情報処理装置１０，２０の各機能が実現される。

【０１１５】
プロセッサＣ１としては、例えば、ＣＰＵ（Central Processing Unit）、ＧＰＵ（Graphic Processing Unit）、ＤＳＰ（Digital Signal Processor）、ＭＰＵ（Micro Processing Unit）、ＦＰＵ（Floating point number Processing Unit）、ＰＰＵ（Physics Processing Unit）、マイクロコントローラ、又は、これらの組み合わせなどを用いることができる。メモリＣ２としては、例えば、フラッシュメモリ、ＨＤＤ（Hard Disk Drive）、ＳＳＤ（Solid State Drive）、又は、これらの組み合わせなどを用いることができる。
【０１１６】
なお、コンピュータＣは、プログラムＰを実行時に展開したり、各種データを一時的に記憶したりするためのＲＡＭ（Random Access Memory）を更に備えていてもよい。また、コンピュータＣは、他の装置との間でデータを送受信するための通信インタフェースを更に備えていてもよい。また、コンピュータＣは、キーボードやマウス、ディスプレイやプリンタなどの入出力機器を接続するための入出力インタフェースを更に備えていてもよい。
【０１１７】
また、プログラムＰは、コンピュータＣが読み取り可能な、一時的でない有形の記録媒体Ｍに記録することができる。このような記録媒体Ｍとしては、例えば、テープ、ディスク、カード、半導体メモリ、又はプログラマブルな論理回路などを用いることができる。コンピュータＣは、このような記録媒体Ｍを介してプログラムＰを取得することができる。また、プログラムＰは、伝送媒体を介して伝送することができる。このような伝送媒体としては、例えば、通信ネットワーク、又は放送波などを用いることができる。コンピュータＣは、このような伝送媒体を介してプログラムＰを取得することもできる。
【０１１８】
〔付記事項１〕
本発明は、上述した実施形態に限定されるものでなく、請求項に示した範囲で種々の変更が可能である。例えば、上述した実施形態に開示された技術的手段を適宜組み合わせて得られる実施形態についても、本発明の技術的範囲に含まれる。
【０１１９】
〔付記事項２〕
上述した実施形態の一部又は全部は、以下のようにも記載され得る。ただし、本発明は、以下の記載する態様に限定されるものではない。
【０１２０】
（付記１）
複数の訓練用例を取得する取得手段と、
用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練する訓練手段と、
前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する選択手段と、
前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する生成手段と、
を備えた情報処理装置。
【０１２１】
上記構成により、予測結果にばらつきがある訓練用例を用いて人工用例を生成するので、決定境界の近傍のように偏った領域に人工用例を生成することがなく、訓練用例が不足する領域に精度よく人工用例を生成することができる。
【０１２２】
（付記２）
前記生成手段は、前記選択した訓練用例と、特徴量空間において前記選択した訓練用例の近傍に存在する用例とを合成して前記人工用例を生成する、付記１に記載の情報処理装

置。
【０１２３】
上記構成により、予測結果にばらつきがある訓練用例の近傍に人工用例を生成するので、訓練用例が不足する領域に精度よく人工用例を生成することができる。
【０１２４】
（付記３）
前記選択手段は、前記複数の訓練用例のうち、２つ以上の前記複数の予測結果にばらつきがある訓練用例を選択し、
前記生成手段は、２つ以上の前記選択した訓練用例を合成して前記人工用例を生成する、付記１に記載の情報処理装置。
【０１２５】
上記構成により、予測結果にばらつきがある訓練用例同士を合成して人工用例を生成するので、訓練用例が不足する領域に精度よく人工用例を生成することができる。
【０１２６】
（付記４）
前記生成手段は、
前記選択した訓練用例と、特徴量空間において前記選択した訓練用例の近傍に存在する用例とを合成する第１生成処理と、
２つ以上の前記選択した訓練用例を合成して前記人工用例を生成する第２生成処理と、
の何れかを実行することにより前記人工用例を生成する、付記１に記載の情報処理装置。
【０１２７】
上記構成により、第１生成処理及び第２生成処理の何れかを選択的に用いて人工用例を生成するので、複数の人工用例を生成する場合には、訓練用例が不足するより多様な領域に人工用例を生成することができる。
【０１２８】
（付記５）
前記人工用例を前記複数の訓練用例に追加して、前記取得手段、前記訓練手段、前記選択手段、及び前記生成手段を再度機能させる、付記１から４の何れか１つに記載の情報処理装置。
【０１２９】
上記構成により、生成した人工用例を加えた訓練用例を用いて機械学習モデル群を訓練することを繰り返すので、予測結果にばらつきがある訓練用例をより精度よく選択することができる。その結果、人工用例がより不足している領域に人工用例を生成することができる。
【０１３０】
（付記６）
前記生成手段は、
複数の前記人工用例を生成し、
複数の前記人工用例のうち類似条件を満たす２つの人工用例を１つの人工用例に統合する、付記１から４の何れか１つに記載の情報処理装置。
【０１３１】
上記構成により統合した人工用例を用いて機械学習モデルを訓練する場合、既に用いた人工用例に類似する人工用例を用いて訓練することを避けられる。したがって、より効率的に機械学習モデルの精度を向上させることができる人工用例を生成することができる。
【０１３２】
（付記７）
前記生成手段は、前記人工用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある人工用例を出力する、付記１から６の何れか１つに記

載の情報処理装置。
【０１３３】
上記構成により出力した人工用例を用いて機械学習モデルを訓練する場合、予測結果のばらつきが小さい人工用例を用いて訓練することを避けられる。したがって、より効率的に機械学習モデルの精度を向上させることができる人工用例を生成することができる。
【０１３４】
（付記８）
前記機械学習モデル群は、前記人工用例を用いて訓練する訓練対象の機械学習モデルを含む、付記１から７の何れか１つに記載の情報処理装置。
【０１３５】
上記構成により、生成した人工用例を用いて訓練対象の機械学習モデルを訓練すれば、より効果的に訓練対象の機械学習モデルの精度を向上させることができる。
【０１３６】
（付記９）
前記機械学習モデル群のうち少なくとも２つは、互いに異なる機械学習アルゴリズムを用いる、付記１から８の何れか１つに記載の情報処理装置。
【０１３７】
上記構成により、予測結果にばらつきがあるより多様な訓練用例を選択することができる。
【０１３８】
（付記１０）
前記機械学習モデル群のそれぞれは、同一の機械学習アルゴリズムを用いる、付記１から８の何れか１つに記載の情報処理装置。
【０１３９】
上記構成により、予測結果にばらつきがある訓練用例をより精度よく選択することができる。
【０１４０】
（付記１１）
前記機械学習モデル群のうち少なくとも１つは決定木である、付記１から１０の何れか１つに記載の情報処理装置。
【０１４１】
上記構成により、決定木の精度をより効果的に向上させることができる人工用例を生成することができる。
【０１４２】
（付記１２）
前記複数の訓練用例及び前記人工用例の一部又は全部にラベルを付与するラベル付与手段をさらに備える、付記１から１１の何れか１つに記載の情報処理装置。
【０１４３】
上記構成により、用例にラベルが付与されていることを前提する訓練手法を用いて、機械学習モデル群又は訓練対象の機械学習モデルを訓練することができる。
【０１４４】
（付記１３）
複数の訓練用例を取得すること、
用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練すること、
前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択すること、及び、
前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成すること、
を含む情報処理方法。

【０１４５】
上記構成により、付記１と同様の効果を奏する。
【０１４６】
（付記１４）
コンピュータを情報処理装置として機能させるためのプログラムであって、前記コンピュータを、
複数の訓練用例を取得する取得手段と、
用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練する訓練手段と、
前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する選択手段と、
前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する生成手段と、
として機能させるプログラム。
【０１４７】
上記構成により、付記１と同様の効果を奏する。
【０１４８】
（付記１５）
付記１４に記載のプログラムが記録された、コンピュータ読み取り可能な記録媒体。
【０１４９】
上記の構成によれば、付記１と同様の効果を奏する。
【０１５０】
〔付記事項３〕
上述した実施形態の一部又は全部は、更に、以下のように表現することもできる。
【０１５１】
少なくとも１つのプロセッサを備え、前記プロセッサは、
複数の訓練用例を取得する取得処理と、
用例を入力として予測結果を出力する機械学習モデルを複数含む機械学習モデル群を、前記複数の訓練用例を用いて訓練する訓練処理と、
前記複数の訓練用例のうち、訓練後の前記機械学習モデル群を用いて得られる複数の予測結果にばらつきがある訓練用例を選択する選択処理と、
前記複数の訓練用例のうち、前記選択した訓練用例を含む２つ以上の訓練用例を合成して人工用例を生成する生成処理と、を実行する情報処理装置。
【０１５２】
なお、この情報処理装置は、更にメモリを備えていてもよく、このメモリには、前記取得処理と、前記訓練処理と、前記選択処理と、前記生成処理とを前記プロセッサに実行させるためのプログラムが記憶されていてもよい。また、このプログラムは、コンピュータ読み取り可能な一時的でない有形の記録媒体に記録されていてもよい。
【符号の説明】
【０１５３】
１０、２０　情報処理装置
１１、２１　取得部
１２、２２　訓練部
１３、２３　選択部
１４、２４　生成部
２５　ラベル付与部
２６　出力部
２７　制御部