(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7424507号
(P7424507)

(45)発行日　令和6年1月30日(2024.1.30)

(24)登録日　令和6年1月22日(2024.1.22)

(51)国際特許分類
G06N 20/00 (2019.01)
G06N 3/096(2023.01)

FI
G06N 20/00
G06N 3/096

請求項の数　8　（全24頁）

(21)出願番号　特願2022-556825(P2022-556825)
(86)(22)出願日　令和2年10月16日(2020.10.16)
(86)国際出願番号　PCT/JP2020/039191
(87)国際公開番号　WO2022/079919
(87)国際公開日　令和4年4月21日(2022.4.21)
審査請求日　令和4年12月21日(2022.12.21)

(73)特許権者　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(74)代理人　110002147
弁理士法人酒井国際特許事務所
(72)発明者　金月 寛彰
神奈川県川崎市中原区上小田中４丁目１番１号 富士通株式会社内
審査官　多賀 実

最終頁に続く

(54)【発明の名称】　検知プログラム、検知方法および検知装置

(57)【特許請求の範囲】

【請求項１】

第１の複数のデータの入力に応じて第１の機械学習モデルから出力された第１の結果と前記第１の複数のデータとに基づいた機械学習によって生成された第２の機械学習モデルに第２の複数のデータを入力し、

前記第２の複数のデータの入力に応じて前記第２の機械学習モデルから出力された第２の結果を取得し、

前記第２の結果と前記第２の機械学習モデルの損失関数の勾配とに基づいて算出される値と閾値との比較に基づいて、前記第１の複数のデータの分布と前記第２の複数のデータの分布との差を検知する、

処理をコンピュータに実行させることを特徴とする検知プログラム。

【請求項２】

前記第１の複数のデータからデータのペアとなる第１ペアデータを選択し、前記第１ペアデータのデータをそれぞれ前記第２の機械学習モデルに入力することで、第１ペアデータのスコア差を算出し、前記スコア差が所定のスコア値未満となるように前記第２の機械学習モデルの出力を調整するハイパーパラメータを算出する処理を更にコンピュータに実行させることを特徴とする請求項１に記載の検知プログラム。

【請求項３】

前記第２の結果を取得する処理は、前記第２の複数のデータからデータのペアとなる第２ペアデータを選択し、前記第２ペアデータのデータをそれぞれ前記第２の機械学習モデ

ルに入力することで、前記第２ペアデータのデータのスコア差を算出し、前記検知する処理は、前記第２ペアデータに関する損失関数の勾配によって、前記第２ペアデータのスコア差を除算した値と、閾値とを基にして、前記第１の複数のデータの分布と前記第２の複数のデータの分布との差を検知することを特徴とする請求項２に記載の検知プログラム。
【請求項４】
前記第１の複数のデータからデータのペアとなる第１ペアデータを選択し、前記第１ペアデータのデータをそれぞれ前記第２の機械学習モデルに入力することで、前記第２ペアデータのデータのスコア差を算出し、前記第１ペアデータに関する損失関数の勾配によって、前記第１ペアデータのスコア差を除算した値を算出する処理を繰り返し実行し、算出した複数の値を基にして前記閾値を算出する処理を更にコンピュータに実行させることを特徴とする請求項３に記載の検知プログラム。
【請求項５】
前記検知する処理によって差が検知されない場合に、前記第２の複数のデータを前記第１の機械学習モデルに入力して、前記第２の複数のデータを分類する処理を更にコンピュータに実行させることを特徴とする請求項１に記載の検知プログラム。
【請求項６】
前記検知する処理によって差が検知された場合に、前記第１の機械学習モデルに対して再度機械学習を実行する処理を更にコンピュータに実行させることを特徴とする請求項１に記載の検知プログラム。
【請求項７】
第１の複数のデータの入力に応じて第１の機械学習モデルから出力された第１の結果と前記第１の複数のデータとに基づいた機械学習によって生成された第２の機械学習モデルに第２の複数のデータを入力し、
前記第２の複数のデータの入力に応じて前記第２の機械学習モデルから出力された第２の結果を取得し、
前記第２の結果と前記第２の機械学習モデルの損失関数の勾配とに基づいて算出される値と閾値との比較に基づいて、前記第１の複数のデータの分布と前記第２の複数のデータの分布との差を検知する、
処理をコンピュータが実行することを特徴とする検知方法。
【請求項８】
第１の複数のデータの入力に応じて第１の機械学習モデルから出力された第１の結果と前記第１の複数のデータとに基づいた機械学習によって生成された第２の機械学習モデルに第２の複数のデータを入力し、前記第２の複数のデータの入力に応じて前記第２の機械学習モデルから出力された第２の結果を取得する取得部と、
前記第２の結果と前記第２の機械学習モデルの損失関数の勾配とに基づいて算出される値と閾値との比較に基づいて、前記第１の複数のデータの分布と前記第２の複数のデータの分布との差を検知する検知部と
を有することを特徴とする検知装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、運用中の機械学習モデルの精度劣化を検知する検知プログラム等に関する。
【背景技術】
【０００２】
近年、企業等で利用されている情報システムに対して、データの判定機能、分類機能等を有する機械学習モデルの導入が進んでいる。以下、情報システムを「システム」と表記する。機械学習モデルは、システム開発時に学習させた訓練データの通りに判定、分類を行うため、システム運用中に入力データの傾向が変化すると、機械学習モデルの精度が劣化する。
【０００３】

図２１は、入力データの傾向の変化による機械学習モデルの劣化を説明するための図である。ここで説明する機械学習モデルは、入力データを第１クラス、第２クラス、第３クラスのいずれかに分類するモデルであり、システム運用前に、訓練データに基づき、予め訓練されているものとする。

【０００４】

図２１において、分布１Ａは、システム運用初期の入力データの分布を示す。分布１Ｂは、システム運用初期からＴ１時間経過した時点の入力データの分布を示す。分布１Ｃは、システム運用初期から更にＴ２時間経過した時点の入力データの分布を示す。時間経過に伴って、入力データの傾向（特徴量等）が変化するものとする。たとえば、入力データが画像であれば、同一の被写体を撮影した画像であっても、季節や時間帯に応じて、入力データの傾向が変化する。

【０００５】

決定境界３は、モデル適用領域３ａ～３ｃの境界を示すものである。たとえば、モデル適用領域３ａは、第１クラスに属する訓練データが分布する領域である。モデル適用領域３ｂは、第２クラスに属する訓練データが分布する領域である。モデル適用領域３ｃは、第３クラスに属する訓練データが分布する領域である。

【０００６】

星印は、第１クラスに属する入力データであり、機械学習モデルに入力した際に、モデル適用領域３ａに分類されることが正しい。三角印は、第２クラスに属する入力データであり、機械学習モデルに入力した際に、モデル適用領域３ｂに分類されることが正しい。丸印は、第３クラスに属する入力データであり、機械学習モデルに入力した際に、モデル適用領域３ａに分類されることが正しい。

【０００７】

分布１Ａでは、全ての入力データが正常なモデル適用領域に分布している。すなわち、星印の入力データがモデル適用領域３ａに位置し、三角印の入力データがモデル適用領域３ｂに位置し、丸印の入力データがモデル適用領域３ｃに位置している。

【０００８】

分布１Ｂでは、入力データの傾向が変化したため、全ての入力データが、正常なモデル適用領域に分布しているものの、星印の入力データの分布がモデル適用領域３ｂの方向に変化している。

【０００９】

分布１Ｃでは、入力データの傾向が更に変化し、星印の一部の入力データが、決定境界３を跨いで、モデル適用領域３ｂに移動しており、適切に分類されておらず、正解率が低下している（機械学習モデルの精度が劣化している）。

【００１０】

ここで、運用中の機械学習モデルの精度劣化を検知する技術として、T^2統計量（Hotelling's T-square）を用いる従来技術がある。この従来技術では、入力データおよび正常データ（訓練データ）のデータ群を主成分分析し、入力データのT^2統計量を算出する。T^2統計量は、標準化した各主成分の原点からデータまでの距離の二乗を合計したものである。従来技術は、入力データ群のT^2統計量の分布の変化を基にして、機械学習モデルの精度劣化を検知する。たとえば、入力データ群のT^2統計量は、異常値データの割合に対応する。

【先行技術文献】

【非特許文献】

【００１１】

【文献】A.Shabbak and H. Midi,"An Improvement of the Hotelling Statistic in Monitoring Multivariate Quality Characteristics",Mathematical Problems in Engineering (2012) 1-15.

【発明の概要】

【発明が解決しようとする課題】

【００１２】
しかしながら、上述した従来技術では、機械学習モデルの精度劣化を引き起こす可能性のあるデータの分布の変化を検知することができないという問題がある。
【００１３】
たとえば、元々の情報量が非常に大きい高次元（数千～数万次元）データでは、従来技術のように、主成分分析により次元を削減すると、ほとんどの情報が失われてしまう。そのため、分類や判定を行うための重要な情報（特徴量）まで落ちてしまい、機械学習モデルの精度劣化を引き起こす可能性のあるデータの分布の変化を検知することができない。
【００１４】
１つの側面では、本発明は、機械学習モデルの精度劣化を引き起こす可能性のあるデータの分布の変化を検知することができる検知プログラム、検知方法および検知装置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】
【００１５】
１つの態様において、検知プログラムは、コンピュータに次の処理を実行させる。コンピュータは、第１の複数のデータの入力に応じて第１の機械学習モデルから出力された第１の結果と第１の複数のデータとに基づいた機械学習によって生成された第２の機械学習モデルに第２の複数のデータを入力する。コンピュータは、第２の複数のデータの入力に応じて第２の機械学習モデルから出力された第２の結果を取得する。コンピュータは、第２の結果と第２の機械学習モデルの損失関数の勾配とに基づいて算出される値と閾値との比較に基づいて、第１の複数のデータの分布と第２の複数のデータの分布との差を検知する。
【発明の効果】
【００１６】
機械学習モデルの精度劣化を引き起こす可能性のあるデータの分布の変化を検知することができる。
【図面の簡単な説明】
【００１７】
【図１】図１は、参考技術を説明するための図である。
【図２】図２は、精度劣化予測の一例を示す図である。
【図３】図３は、コンセプトドリフトの一例を示す図である。
【図４】図４は、インスペクターモデルの基本的な仕組みを説明するための図である。
【図５】図５は、参考技術の問題を説明するための図である。
【図６】図６は、統計的検定の問題を説明するための図である。
【図７】図７は、本実施形態に係る検知装置の処理を説明するための図である。
【図８】図８は、知識蒸留を説明するための図である。
【図９】図９は、本実施形態に係る検知装置の構成を示す機能ブロック図である。
【図１０】図１０は、訓練データセットのデータ構造の一例を示す図である。
【図１１】図１１は、運用モデルの一例を説明するための図である。
【図１２】図１２は、疑似サンプルテーブルのデータ構造の一例を示す図である。
【図１３】図１３は、蒸留データテーブルのデータ構造の一例を示す図である。
【図１４】図１４は、運用データセットテーブルのデータ構造の一例を示す図である。
【図１５】図１５は、本実施形態に係る特徴空間の決定境界を説明するための図である。
【図１６】図１６は、ハイパーパラメータに応じたスコア差の分布を示す図である。
【図１７】図１７は、本実施形態に係る検知装置の処理手順を示すフローチャートである。
【図１８】図１８は、各機械学習モデルの決定境界の性質を示す図（１）である。
【図１９】図１９は、各機械学習モデルの決定境界の性質を示す図（２）である。
【図２０】図２０は、本実施形態に係る検知装置と同様の機能を実現するコンピュータのハードウェア構成の一例を示す図である。
【図２１】図２１は、入力データの傾向の変化による機械学習モデルの劣化を説明するた

めの図である。

【発明を実施するための形態】

【００１８】

以下に、本願の開示する検知プログラム、検知方法および検知装置の実施例を図面に基づいて詳細に説明する。なお、この実施例によりこの発明が限定されるものではない。

【００１９】

本実施形態の説明を行う前に、機械学習モデルの精度劣化を検知する参考技術について説明する。参考技術では、異なる条件でモデル適用領域を狭めた複数の監視器を用いて、機械学習モデルの精度劣化を検知する。以下の説明では、監視器を「インスペクターモデル」と表記する。

【００２０】

図１は、参考技術を説明するための図である。機械学習モデル１０は、訓練データを用いて機械学習を実行することで生成される機械学習モデルである。参考技術では、機械学習モデル１０の精度劣化を検知する。たとえば、訓練データは、機械学習モデル１０のパラメータを訓練する場合に用いられるものであり、正解ラベルが対応付けられる。

【００２１】

インスペクターモデル１１Ａ，１１Ｂ，１１Ｃは、それぞれ異なる条件でモデル適用領域が狭められ、異なる決定境界を有する。参考技術では、訓練データに何らかの改変を加え、改変を加えた訓練データを用いて、インスペクターモデル１１Ａ～１１Ｃを訓練している。

【００２２】

インスペクターモデル１１Ａ～１１Ｃは、それぞれ決定境界が異なるため、同一の入力データを入力しても、出力結果が異なる場合がある。参考技術では、インスペクターモデル１１Ａ～１１Ｃの出力結果の違いを基にして、機械学習モデル１０の精度劣化を検知する。図１に示す例では、インスペクターモデル１１Ａ～１１Ｃを示すが、他のインスペクターモデルを用いて、精度劣化を検知してもよい。インスペクターモデル１１Ａ～１１ＣにはＤＮＮ（Deep Neural Network）を利用する。

【００２３】

参考技術では、インスペクターモデル１１Ａ～１１Ｃの出力結果が全て同じである場合に、機械学習モデル１０の精度が劣化していないと判定する。一方、参考技術では、インスペクターモデル１１Ａ～１１Ｃの出力結果が異なる場合に、機械学習モデル１０の精度劣化を検知する。

【００２４】

図２は、精度劣化予測の一例を示す図である。図２のグラフの縦軸は、精度に対応する軸であり、横軸は時刻に対応する軸である。図２に示すように、時間経過に伴って、精度が低下しており、時刻ｔ１において、精度の許容限界を下回る。たとえば、参考技術では、時刻ｔ１において、精度劣化（許容限界を下回ったこと）を検知する。

【００２５】

時間経過に伴う入力データの分布（特徴量）の変化をコンセプトドリフトと呼ぶ。図３は、コンセプトドリフトの一例を示す図である。図３の縦軸は、第１の特徴量に対応する軸であり、横軸は、第２の特徴量に対応する軸である。たとえば、機械学習モデル１０の運用開始時において、第１クラスに対応する第１データの分布を分布$Ａ_1$とし、第２クラスに対応する第２データの分布を分布Ｂとする。

【００２６】

時間経過に伴って、第１データの分布$Ａ_1$が、分布$Ａ_2$に変化する場合がある。オリジナルの機械学習モデル１０は、第１データの分布を、分布$Ａ_1$として学習を行っているため、時間経過に伴って精度が下がり、再訓練が必要となる。

【００２７】

コンセプトドリフトが発生するデータには、スパムメール、電気需要予測、株価予測、ポーカーハンドの戦略手順、画像等が含まれる。たとえば、画像は、季節や時間帯によっ

て、同一の被写体であっても、画像の特徴量が異なる。

【００２８】

ここで、上述した参考技術では、機械学習モデル１０の精度劣化を検知するために、複数のインスペクターモデル１１Ａ～１１Ｃを訓練している。そして、複数のインスペクターモデル１１Ａ～１１Ｃを訓練するためには、機械学習モデル１０や、機械学習モデル１０の訓練時に用いた、訓練データに何らかの改変を加えることができるという条件が必須である。たとえば、機械学習モデル１０が確信度を算出するモデルであること等、機械学習モデル１０が特定の機械学習モデルであることが求められる。

【００２９】

図４は、インスペクターモデルの基本的な仕組みを説明するための図である。たとえば、インスペクターモデルは、第１クラスに属する訓練データの分布A_1と、第２クラスに属する訓練データの分布Ｂとの境界となる決定境界５の機械学習を実行することで、作成される。時間経過に伴う、運用データに対する機械学習モデル１０の精度劣化を検出するためには、決定境界５の危険領域５ａを監視し、危険領域５ａに含まれる運用データの数が増加したか（または減少したか）否かを特定し、運用データの数が増加した（または減少した）場合に、精度劣化を検知する。

【００３０】

ここで、上述した参考技術の問題について説明する。図５は、参考技術の問題を説明するための図である。たとえば、インスペクターモデル等を用いて、危険領域５ａを監視する手法を用いるとケース１またはケース２のようになる。

【００３１】

ケース１のように、時間経過に伴って、データの分布Ａと、データの分布Ｂとが近づく方向に変化した場合に、危険領域５ａのデータの数が変化するため、参考技術によって、精度劣化を検知することができる。

【００３２】

一方、ケース２のように、時間経過に伴って、データの分布Ａと、データの分布Ｂとが離れていく方向に変化した場合、危険領域５ａに含まれるデータの数が変化しないため、参考技術によって、精度劣化を検知することができない。

【００３３】

ケース２のように、データの分布Ａと、データの分布Ｂとが離れていく場合、データの答えそのものが変化している可能性があるが、参考技術で説明したインスペクターモデルを用いても、各インスペクターモデルから出力される結果に変化はない。たとえば、インスペクターモデルは、データが、第１クラスの領域に含まれていれば、決定境界からいくら離れても、データを第１クラスに分類する。参考技術では、データの答えが変化していないことを暗黙的に仮定している。

【００３４】

ところで、決定境界によらないデータのシフトは統計的検定で検出するが可能である。たとえば、統計的検定には、Student検定、Kolmogorov Smirnov検定、Ｌ２距離を用いる方法、コサイン距離を用いる方法、ＫＬ距離を用いる方法、ワーサースタイン距離を用いる方法等がある。

【００３５】

しかし、統計的検定を用いると、分類に関係のない特徴量（ノイズ）の移動もすべて検知してしまうため、誤検知が多いという問題がある。図６は、統計的検定の問題を説明するための図である。図６では、ｘ軸、ｙ軸、ｚ軸からなる特徴空間に位置するデータ群６ａと、６ａ－１を用いて説明する。ここで、決定境界７が、ｘｙ平面に位置していると、ｚ軸方向の変化は、分類結果と無関係となる。しかし、統計的検定を用いると、時間経過に伴って、データ群６ｂ－１がｚ軸方向に移動し、データ群６ｂ－２に変化すると、かかる変化を検知し、誤検知となる。

【００３６】

次に、本実施形態に係る検知装置の処理の一例について説明する。図７は、本実施形態

に係る検知装置の処理を説明するための図である。図７では、学習フェーズおよび運用フェーズにおける検知装置の処理について説明する。

【００３７】

検知装置の学習フェーズの処理について説明する。検知装置は、訓練データセット１４１を入力として、運用モデル５０の機械学習を実行する。訓練データセット１４１は、複数の訓練データを含み、複数の訓練データには、正解ラベルが付与される。検知装置は、機械学習を実行済みの運用モデル５０に訓練データセット１４１を入力した場合の出力と、訓練データセット１４１とを入力として、インスペクターモデル５５の機械学習を実行する。たとえば、検知装置は、知識蒸留（ＫＤ：Knowledge Distiller）によって、インスペクターモデル５５の機械学習を実行する。

【００３８】

図８は、知識蒸留を説明するための図である。知識蒸留では、Teacherモデル７Ａの出力値を模倣するような、Studentモデル７Ｂを構築する。Teacherモデル７Ａは、図７の運用モデル５０に対応する。Studentモデル７Ｂは、図７のインスペクターモデル５５に対応する。たとえば、訓練データ６が与えられ、訓練データ６には正解ラベル「犬」が付与されているものとする。説明の便宜上、Teacherモデル７AおよびStudentモデル７ＢをＮＮ（Neural Network）とするが、これに限定されるものではない。

【００３９】

検知装置は、訓練データ６を入力した際のTeacherモデル７Ａの出力結果が、正解ラベル「犬」に近づくように、Teacherモデル７Ａのパラメータを訓練（誤差逆伝播法による機械学習を実行）する。また、検知装置は、訓練データ６を入力した際のStudentモデル７Ｂの出力結果が、訓練データ６を入力した際のTeacherモデル７Ａの出力結果に近づくように、Studentモデル７Ｂのパラメータを訓練する。Teacherモデル７Ａの出力を「ソフトターゲット（Soft Target）」と呼ぶ。訓練データの正解ラベルを「ハードターゲット（Hard Target）」と呼ぶ。

【００４０】

上記のように、Teacherモデル７Ａを、訓練データ６とハードターゲットとを用いて訓練し、Studentモデル７Ｂを、訓練データ６とソフトターゲットとを用いて訓練する手法を、知識蒸留と呼ぶ。検知装置は、他の訓練データについても同様にして、Teacherモデル７AおよびStudentモデル７Ｂを訓練する。たとえば、検知装置は、訓練データセット１４１と、運用モデル５０から出力されるソフトターゲットとを用いて、インスペクターモデル５５の機械学習を実行する。

【００４１】

図７の説明に戻る。検知装置の運用フェーズの処理について説明する。検知装置は、運用データセットＣ０に含まれる複数の運用データをインスペクターモデル５５に入力し、インスペクターモデル５５から出力される結果を取得する。検知装置は、インスペクターモデル５５から出力される結果およびインスペクターモデル５５の損失係数の勾配を基に算出される値と、閾値とを比較し、コンセプトドリフトを検知する。

【００４２】

たとえば、インスペクターモデル５５から出力される結果およびインスペクターモデル５５の損失係数の勾配を基に算出される値は、モデル適用領域の境界を示す決定境界からの距離を示す。以下の説明では、インスペクターモデル５５から出力される結果およびインスペクターモデル５５の損失係数の勾配を基に算出される値を「評価値」と表記する。評価値が、閾値以上である場合には、インスペクターモデル５５に入力した運用データセットが、決定境界から大きく離れており、コンセプトドリフトが発生していることを意味する。検知装置は、評価値が閾値以上となった場合に、コンセプトドリフトを検知する。

【００４３】

検知装置は、ドリフトを検知していない場合には、運用データセットＣ０を、運用モデル５０に入力して、運用データセットＣ０のデータの属するクラスを予測する。一方、検知装置は、ドリフトを検知した場合には、新たな訓練データセットによって、運用モデル

５０に対して機械学習を再度実行する。

【００４４】

上記のように、本実施形態に係る検知装置は、知識蒸留を用いて、運用モデル５０の監視器となるインスペクターモデル５５の機械学習を実行する。検知装置は、インスペクターモデル５５に、運用データセットＣ０を入力し、インスペクターモデル５５から出力される結果およびインスペクターモデル５５の損失係数の勾配を基に算出される評価値と、閾値とを比較し、コンセプトドリフトを検知する。これによって、時間経過に伴い、データの分布が、決定境界から離れる方向に変化しても、かかるデータの分布の変化を検知することができ、運用モデル５０の精度劣化を検知することができる。

【００４５】

次に、本実施形態に係る検知装置の構成の一例について説明する。図９は、本実施形態に係る検知装置の構成を示す機能ブロック図である。図９に示すように、この検知装置１００は、通信部１１０、入力部１２０、出力部１３０、記憶部１４０、制御部１５０を有する。

【００４６】

通信部１１０は、ネットワークを介して、外部装置（図示略）とデータ通信を実行する。たとえば、通信部１１０は、外部装置から後述する訓練データセット１４１等を受信する。

【００４７】

入力部１２０は、データを入力するための装置またはインタフェースである。たとえば、入力部１２０は、マウス、およびキーボードなどである。

【００４８】

出力部１３０は、画面を表示するディスプレイなどである。

【００４９】

記憶部１４０は、データや制御部１５０が実行するプログラムなどを記憶する記憶装置の一例であり、たとえば、ハードディスクやメモリなどである。記憶部１４０は、訓練データセット１４１と、運用モデルデータ１４２、疑似サンプルテーブル１４３、蒸留データテーブル１４４、インスペクターモデルデータ１４５、運用データセットテーブル１４６とを有する。

【００５０】

訓練データセット１４１は、複数の訓練データを含む。図１０は、訓練データセットのデータ構造の一例を示す図である。図１０に示すように、この訓練データセットは、レコード番号と、訓練データと、正解ラベルとを対応付ける。レコード番号は、訓練データと、正解ラベルとの組を識別する番号である。訓練データは、メールスパムのデータ、電気需要予測、株価予測、ポーカーハンドのデータ、画像データ等に対応し、複数の次元の特徴量を含む。正解ラベルは、第１クラスまたは第２クラスを一意に識別する情報である。

【００５１】

運用モデルデータ１４２は、運用モデル５０（機械学習モデル）のデータである。本実施形態の運用モデル５０は、所定の分類アルゴリズムによって、入力データを、複数のクラスに分類する。本実施形態では、運用モデル５０を、ＮＮとして説明を行う。

【００５２】

図１１は、運用モデルの一例を説明するための図である。図１１に示すように、運用モデル５０は、ニューラルネットワークの構造を有し、入力層５０ａ、隠れ層５０ｂ、出力層５０ｃを持つ。入力層５０ａ、隠れ層５０ｂ、出力層５０ｃは、複数のノードがエッジで結ばれる構造となっている。隠れ層５０ｂ、出力層５０ｃは、活性化関数と呼ばれる関数とバイアス値とを持ち、エッジには、重みが設定される。以下の説明では、バイアス値および重みを「パラメータ」と表記する。

【００５３】

入力層５０ａに含まれる各ノードに、データ（データの特徴量）を入力すると、隠れ層２０ｂを通って、出力層２０ｃから、各クラスの確率が出力される。

【００５４】
疑似サンプルテーブル１４３は、訓練データセット１４１を基に生成される複数の疑似サンプルを保持する。図１２は、疑似サンプルテーブルのデータ構造の一例を示す図である。図１２に示すように、この疑似サンプルテーブル１４３は、サンプル番号と、疑似サンプルとを対応付ける。サンプル番号は、疑似サンプルを識別する情報である。疑似サンプルは、訓練データの特徴量をスケーリングしたデータである。
【００５５】
蒸留データテーブル１４４は、疑似サンプルテーブル１４３の各疑似サンプルを、運用モデル５０に入力した場合の出力結果（ソフトターゲット）を格納する。図１３は、蒸留データテーブルのデータ構造の一例を示す図である。図１３に示すように、この蒸留データテーブル１４４は、サンプル番号と、疑似サンプルと、ソフトターゲットとを対応付ける。サンプル番号、疑似サンプルの説明は、図１２で行ったサンプル番号、疑似サンプルの説明と同様である。ソフトターゲットは、疑似サンプルを運用モデル５０に入力した場合の出力結果である。たとえば、ソフトターゲットは、複数のクラスのうち、いずれかのクラスとなる。
【００５６】
インスペクターモデルデータ１４５は、インスペクターモデル５５のデータである。インスペクターモデル５５は、図１１で説明した運用モデル５０と同様にして、ニューラルネットワークの構造を有し、入力層、隠れ層、出力層を持つ。インスペクターモデル５５には、パラメータが設定される。インスペクターモデル５５のパラメータは、知識蒸留によって訓練される。
【００５７】
運用データセットテーブル１４６は、時間経過に伴って、追加される運用データセットを有する。図１４は、運用データセットテーブルのデータ構造の一例を示す図である。図１４に示すように、運用データセットテーブル１４６は、データ識別情報と、運用データセットとを有する。データ識別情報は、運用データセットを識別する情報である。運用データセットは、複数の運用データを含む。運用データは、メールスパムのデータ、電気需要予測、株価予測、ポーカーハンドのデータ、画像データ等に対応する。
【００５８】
図９の説明に戻る。制御部１５０は、検知装置１００全体を司る処理部であり、生成部１５１、算出部１５２、取得部１５３、検知部１５４を有する。制御部１５０は、たとえば、プロセッサ等である。
【００５９】
生成部１５１は、運用モデルデータ１４２を生成する処理、疑似サンプルテーブル１４３を生成する処理、蒸留データテーブル１４４を生成する処理、インスペクターモデルデータ１４５を生成する処理を実行する。
【００６０】
生成部１５１が、運用モデルデータ１４２を生成する処理について説明する。生成部１５１は、訓練データセット１４１を入力として、運用モデル５０の機械学習を実行する。たとえば、生成部１５１は、訓練データセットの訓練データを、運用モデル５０の入力層に入力した場合、出力層の出力結果が、入力した訓練データの正解ラベルに近づくように、運用モデル５０のパラメータを訓練する。たとえば、生成部１５１は、誤差逆伝播法による機械学習を実行する。生成部１５１は、機械学習を実行した運用モデル５０のデータ（運用モデルデータ１４２）を、記憶部１４０に登録する。
【００６１】
図１５は、本実施形態に係る特徴空間の決定境界を説明するための図である。特徴空間３０は、訓練データセット１４１の各訓練データを可視化したものある。特徴空間３０の横軸は、第１特徴量の軸に対応し、縦軸は、第２特徴量の軸に対応する。ここでは説明の便宜上、２軸で各訓練データを示すが、訓練データは、多次元のデータであるものとする。たとえば、丸印の訓練データに対応する正解ラベルを「第１クラス」とし、三角印の訓

練データに対応する正解ラベルを「第２クラス」とする。

【００６２】

たとえば、訓練データセット１４１によって、運用モデル５０を学習すると、特徴空間３０は、決定境界３１によって、モデル適用領域３１Ａと、モデル適用領域３１Ｂとに分類される。たとえば、運用モデル５０が、ＮＮである場合、運用モデル５０に運用データを入力すると、第１クラスの確率と、第２クラスの確率とが出力される。第１クラスの確率が、第２クラスよりも大きい場合には、データは、第１クラスに分類される。第２クラスの確率が、第１クラスよりも大きい場合には、データは、第２クラスに分類される。

【００６３】

図１５では、訓練データの正解ラベルが「第１クラス」または「第２クラス」となる場合について説明するが、他のクラスの正解ラベルが付与されてもよい。ｎ種類のクラスが存在する場合には、特徴空間３０には、ｎ個のモデル適用領域が設定される。運用モデル５０が、ＮＮである場合、運用モデル５０に運用データを入力すると、各クラスの確率が出力される。

【００６４】

生成部１５１が、疑似サンプルテーブル１４３を生成する処理について説明する。生成部１５１は、訓練データセット１４１に含まれる各訓練データに対してデータ変換を実行する。たとえば、生成部１５１は、訓練データの各次元の特徴量の値が、０以上、１未満の値に含まれるようなデータ変換（Min-Max Scaling）を実行する。

【００６５】

生成部１５１は、データ変換後の各訓練データのうち、各次元の特徴量の値が、－ｍ以上、１＋ｍ未満となる訓練データをランダムに選択する。「ｍ」は、marginであり、あらかじめ任意の実数が設定される。上記処理によって生成部１５１がランダムに選択したデータ変換後の訓練データを「疑似サンプル」と表記する。たとえば、疑似サンプルの特徴量の値の範囲は、式（１）によって定義される。特徴量の次元数をｎとする。

【００６６】

【数１】

$$[0,1] \in \mathbb{R}^n \quad \cdots(1)$$

【００６７】

生成部１５１は、サンプル番号と、疑似サンプルと、正解ラベルとを対応付けて、疑似サンプルテーブル１４３に登録する。疑似サンプルの正解ラベルは、疑似サンプルに対応するデータ変換前の訓練データの正解ラベルとする。

【００６８】

生成部１５１が、蒸留データテーブル１４４を生成する処理について説明する。生成部１５１は、疑似サンプルテーブル１４３の疑似サンプルを、運用モデル５０に入力し、運用モデル５０の出力結果（ソフトターゲット）を取得する。生成部１５１は、サンプル番号、疑似サンプル、ソフトターゲットを、蒸留データテーブル１４４に登録する。

【００６９】

生成部１５１は、疑似サンプルテーブル１４３の各疑似サンプルについて、上記処理を繰り返し実行することで、ソフトターゲットを取得し、蒸留データテーブル１４４に登録する。

【００７０】

たとえば、蒸留データテーブル１４４に登録される疑似サンプルと、ソフトターゲットとの組の集合を、疑似データ集合とすると、疑似データ集合は、式（２）によって定義される。式（２）の位置ａ１に示す記号を「Ｄハット」と表記する。式（２）のａ２に示す記号を「ｘハット」と表記する。Ｄハットは、疑似データ集合を示す。ｘハットは、疑似サンプルを示す。ｆ（ｘハット）は、運用モデル５０から出力されるソフトターゲットである。式（２）の位置ａ３に示す記号を「花文字Ｘ」と表記する。花文字Ｘは、入力空間

を示す。

【００７１】

【数２】

$$\underset{a1}{\widehat{\mathcal{D}}} = \{\underset{a2}{\hat{x}}, f(\hat{x})\} \qquad (\forall \hat{x} \in \underset{a3}{\mathcal{X}}; \mathcal{X} \in \mathbb{R}^d) \qquad \cdots (2)$$

【００７２】

生成部１５１が、インスペクターモデルデータ１４５を生成する処理について説明する。生成部１５１は、蒸留データテーブル１４４を取得し、蒸留データテーブル１４４を基にして、インスペクターモデル５５のパラメータを訓練する。たとえば、生成部１５１は、蒸留データテーブル１４４の疑似サンプルを、インスペクターモデル５５の入力層に入力した場合に、出力層の出力結果が、ソフトターゲットに近づくように、インスペクターモデル５５のパラメータを訓練する。たとえば、生成部１５１は、誤差逆伝播法による機械学習を実行する。生成部１５１は、機械学習を実行したインスペクターモデル５５のデータ（インスペクターモデルデータ１４５）を、記憶部１４０に登録する。

【００７３】

上記の生成部１５１が、インスペクターモデル５５を訓練する処理は、式（３）に示す損失関数$\theta_2{}^*$を最小化させるように、インスペクターモデル５５を訓練することに対応する。式（３）において、ｆ（Ｘ；θ_1）は、運用モデル５０の出力に対応し、Ｘは、Ｄハットに対応する。θ_1は、運用モデル５０のパラメータを示し、訓練済みのパラメータとなる。ｇ（Ｘ；θ_2）は、インスペクターモデル５５の出力に対応し、Ｘは、Ｄハットに対応する。θ_2は、インスペクターモデル５５のパラメータを示し、訓練対象のパラメータとなる。

【００７４】

【数３】

$$\theta_2^* = \underset{\theta_2}{argmin}\, \mathcal{L}\left(f\left(X;\theta_1\right), g\left(X;\theta_2\right)\right) \qquad \cdots (3)$$

【００７５】

生成部１５１は、後述する検知部１５４から、ドリフトを検知した旨を受け付けた場合には、運用モデル５０およびインスペクターモデル５５の機械学習を再度実行する。たとえば、生成部１５１は、最新の訓練データセット１４１を、外部装置から取得し、最新の訓練データセット１４１を用いて、運用モデル５０およびインスペクターモデル５５を再度訓練する。

【００７６】

算出部１５２は、温度付きソフトマックス（softmax）を用いて、インスペクターモデル５５の出力をスケールするためのハイパーパラメータを算出する。データｉを入力した場合の、温度付きソフトマックスを用いたインスペクターモデル５５の出力g_iは、式（４）によって定義される。式（４）において、「z_i」は、データｉを入力した際の、インスペクターモデル５５の出力であって、通常のソフトマックスを用いたインスペクターモデル５５の出力を示す。「Ｔ」は、ハイパーパラメータを示す。以下の算出部１５２の説明では、温度付きソフトマックスを用いたインスペクターモデル５５の出力を、スコアと表記する。

【００７７】

【数４】

$$g_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \qquad \cdots (4)$$

【００７８】
算出部１５２は、疑似サンプルテーブル１４３から、疑似サンプルのペアを選択する。ここでは、疑似サンプルのペアを、第１疑似サンプル、第２疑似サンプルとする。算出部１５２は、第１疑似サンプルを、温度付きソフトマックスを用いたインスペクターモデル５５に入力して、第１スコアを算出する。算出部１５２は、第２疑似サンプルを、温度付きソフトマックスを用いたインスペクターモデル５５に入力して、第２スコアを算出する。算出部１５２は、第１スコアと第２スコアとの差分の絶対値を、スコア差として算出する。
【００７９】
算出部１５２は、疑似サンプルテーブル１４３から、異なる疑似サンプルのペアを選択し、スコア差を算出する処理を繰り返し実行する。算出部１５２は、複数のスコア差のうち、最大のスコア差が閾値Ｔｈｓ未満となるような、ハイパーパラメータを探索する。閾値Ｔｈｓは、あらかじめ設定される。
【００８０】
図１６は、ハイパーパラメータに応じたスコア差の分布を示す図である。図１６において、グラフＧ１－１は、ハイパーパラメータがＴ＝１の場合の、スコア差の頻度を示すヒストグラムである。グラフＧ１－１の横軸はスコア差に対応し、グラフＧ１－１の縦軸は頻度に対応する。グラフＧ１－１に示す例では、スコア差０～０．１において、わずかに頻度があり、スコア差０．９～１．０において、頻度が集中している。
【００８１】
グラフＧ１－２は、ハイパーパラメータがＴ＝１の場合の、決定境界からの距離を色によって表現している。グラフＧ１－２上の色が、バーｂａの下方向の色に近づくほど、決定境界からの距離が近いことを示す。グラフＧ１－２上の色が、バーｂａの上方向の色に近づくほど、決定境界からの距離が遠いことを示す。グラフＧ１－２では、色がほぼ二極化しており、決定境界からの距離を細かく分類できていない。
【００８２】
グラフＧ２－１は、ハイパーパラメータがＴ＝１００００の場合の、スコア差の頻度を示すヒストグラムである。グラフＧ２－１の横軸はスコア差に対応し、グラフＧ２－１の縦軸は頻度に対応する。グラフＧ２－１に示す例では、グラフＧ１－１と比較して、スコア差の頻度がまんべんなく分布している。
【００８３】
グラフＧ２－２は、ハイパーパラメータがＴ＝１００００の場合の、決定境界からの距離を色によって表現している。グラフＧ２－２上の色が、バーｂａの下方向の色に近づくほど、決定境界からの距離が近いことを示す。グラフＧ２－２上の色が、バーｂａの上方向の色に近づくほど、決定境界からの距離が遠いことを示す。グラフＧ２－２では、グラフＧ２－１と比較して、決定境界からの距離を細かく分類できている。
【００８４】
算出部１５２が、最大のスコア差が閾値Ｔｈｓ未満となるようなハイパーパラメータを探索すると、スコア差と頻度との関係は、グラフＧ２－１の関係に近づき、グラフＧ２－２で示すように、決定境界からの距離を細かく分類できるようになる。算出部１５２は、算出（探索）したハイパーパラメータの情報を、検知部１５４に出力する。
【００８５】
検知部１５４は、インスペクターモデル５５に運用データを入力して、決定境界からの距離を算出し、決定境界からの距離を基にして、訓練データセット１４１の分布と、運用データセットの分布との差を検知する。検知部１５４は、訓練データセット１４１の分布と、運用データセットの分布との差を、ドリフトとして検知し、ドリフトを検知した旨を、生成部１５１に出力する。
【００８６】
決定境界からの距離は、式（５）によって近似することができる。式（５）の位置ｂ１に示す記号を単に「ｄ～」と表記する。

【００８７】
【数５】

$$\tilde{d}_{\substack{g,x,t,\{i,j\}\\ \uparrow\\ b1}} = \frac{\left|g_i(x_t) - g_j(x_t)\right|}{\left\|\nabla_x g_i(x_t) - \nabla_x g_j(x_t)\right\|_q} \quad \cdots(5)$$

【００８８】
検知部１５４は、運用データセットから、運用データのペアとなる運用データｉと、運用データｊとを選択するものとする。式（５）において、ｇｉ（ｘｔ）は、運用データｉを入力することで「温度付きソフトマックスを用いたインスペクターモデル５５」から出力される出力結果である。ｇｊ（ｘｔ）は、運用データｊを入力することで「温度付きソフトマックスを用いたインスペクターモデル５５」から出力される出力結果である。
【００８９】
∇ｘｇｉ（ｘｔ）は、偏微分した「温度付きソフトマックスを用いたインスペクターモデル５５」に運用データｉを入力した際の出力結果であり、運用データｉにおける損失関数の勾配に対応する。∇ｘｇｊ（ｘｔ）は、偏微分した「温度付きソフトマックスを用いたインスペクターモデル５５」に運用データｊを入力した際の出力結果であり、運用データｊにおける損失関数の勾配に対応する。
【００９０】
式（５）の分母は、運用データｉにおける損失関数の勾配と、運用データｊにおける損失関数の勾配との差分のｑ－ｎｏｒｍを示す。ｑ－ｎｏｒｍは、ｐ－ｎｏｒｍのｄｕａｌ－ｎｏｒｍであり、ｐとｑとは式（６）の関係が成り立つ。たとえば、式（６）により、ｑ＝１ならｐ＝∞、ｑ＝２ならｐ＝２、ｑ＝∞ならｐ＝１となる。
【００９１】
１／ｐ＋１／ｑ＝１・・・（６）
【００９２】
ｐ－ｎｏｒｍは、式（７）によって示される。
【００９３】
【数６】

$$\|x\|_p = \sqrt[p]{\sum |x_i|^p} \quad \cdots(7)$$

【００９４】
たとえば、ｑ＝１を計算する場合には、ｐ＝∞を計算すればよく、無限大ノルムは、式（８）によって示される。なお、１－ｎｏｒｍは、式（９）によって示され、２－ｎｏｒｍは、式（１０）によって示される。
【００９５】
【数７】

$$\|x\|_\infty = \max_i |x_i| \quad \cdots(8)$$

【数８】

$$\|x\|_1 = \sum |x_i| \quad \cdots(9)$$

【数９】

$$\|x\|_2 = \sqrt{\sum |x_i|^2} \quad \cdots(10)$$

【００９６】
検知部１５４は、運用データセットから、運用データのペアを再度選択し、再度選択した運用データのペアと、式（５）とを基にして、ｄ～を算出する処理を繰り返し実行する。検知部１５４は、算出した複数のｄ～の平均値を算出する。複数のｄ～の平均値は、上述した評価値に対応する。検知部１５４は、評価値が、閾値ｔｈσ以上となった場合に、訓練データセット１４１の分布と、運用データセットの分布との差を検知し、ドリフトとして検知する。検知部１５４は、ドリフトを検知した旨を、外部装置に通知してもよい。
【００９７】
検知部１５４は、運用データセットテーブル１４６に複数の運用データセットが登録されている場合には、各運用データセットについて、上記処理を繰り返し実行する。
【００９８】
ところで、検知部１５４は、次の処理を実行して、閾値ｔｈσを算出しておく。検知部１５４は、疑似サンプルテーブル１４３から、疑似サンプルのペアを選択し、選択した疑似サンプルのペアと、式（５）とを基にして、ｄ～を算出する処理を繰り返し実行する。検知部１５４は、算出した複数のｄ～を基にして、ｄ～の標準偏差を算出し、算出した標準偏差を、上記の閾値Ｔｈσとして設定する。
【００９９】
分類部１５５は、運用データセットの運用データを、運用モデル５０に入力することで、運用データの属するクラスを特定する。分類部１５５は、運用データセットの他の運用データに対して、上記処理を繰り返し実行することで、複数の運用データを複数のクラスに分類する。
【０１００】
次に、本実施形態に係る検知装置１００の処理手順の一例について説明する。図１７は、本実施形態に係る検知装置の処理手順を示すフローチャートである。たとえば、検知装置１００は、運用データセットテーブル１４６に新たな運用データセットが登録されるたびに、図１７の処理を繰り返し実行する。
【０１０１】
図１７に示すように、検知装置１００の生成部１５１は、訓練データセット１４１を入力として、運用モデルの機械学習を実行する（ステップＳ１０１）。
【０１０２】
生成部１５１は、訓練データセット１４１を基にして、疑似サンプルテーブル１４３を生成する（ステップＳ１０２）。生成部１５１は、疑似サンプルテーブル１４３の疑似サンプルを運用モデルに入力することで、蒸留データテーブル１４４を生成する（ステップＳ１０３）。
【０１０３】
生成部１５１は、知識蒸留によって、運用モデル５０を模倣するインスペクターモデル５５の機械学習を実行する（ステップＳ１０４）。検知装置１００の算出部１５２は、インスペクターモデル５５のハイパーパラメータを算出する（ステップＳ１０５）。
【０１０４】
検知装置１００の取得部１５３は、運用データセットの運用データのペアをインスペクターモデル５５に入力し、インスペクターモデルの出力結果を取得する（ステップＳ１０６）。検知装置１００の検知部１５４は、式（５）を基にして、ドリフトを検知したか否かを判定する（ステップＳ１０７）。
【０１０５】
検知部１５４は、ドリフトを検知した場合には（ステップＳ１０８，Ｙｅｓ）、ステップＳ１０１に移行する。検知部１５４は、ドリフトを検知していない場合には（ステップＳ１０８，Ｎｏ）、ステップＳ１０９に移行する。検知装置１００の分類部１５５は、運用データを運用モデル５０に入力し、運用データをクラスに分類する（ステップＳ１０９）。
【０１０６】

次に、本実施形態に係る検知装置１００の効果について説明する。検知装置１００は、知識蒸留を用いて、運用モデル５０の監視器となるインスペクターモデル５５の機械学習を実行する。検知装置１００は、インスペクターモデル５５に、運用データセットを入力し、インスペクターモデル５５から出力される結果およびインスペクターモデル５５の損失係数の勾配を基に算出される評価値と、閾値とを比較し、コンセプトドリフトを検知する。これによって、時間経過に伴い、データの分布が、決定境界から離れる方向に変化しても、かかるデータの分布の変化を検知することができ、運用モデル５０の精度劣化を検知することができる。

【０１０７】

検知装置１００は、異なる疑似サンプルのペアを選択し、スコア差を算出する処理を繰り返し実行する。検知装置１００は、複数のスコア差のうち、最大のスコア差が閾値Ｔｈｓ未満となるような、ハイパーパラメータを算出する。これによって、決定境界からの距離を、段階的に数値化することができる。

【０１０８】

検知装置１００は、式（５）を基にして、評価値を算出して、閾値Ｔｈσとの比較を行い、データの分布の変化を検知する。また、検知装置１００は、訓練データセットを用いて、閾値Ｔｈσを算出しておく。これによって、ドリフトを精度よく検知することができる。

【０１０９】

検知装置１００は、ドリフトを検知しない場合に、運用データセットの運用データを運用モデル５０に入力して、運用データを複数のクラスに分類する。このため、ドリフト発生前の運用モデル５０によって、運用データを適切に、複数のクラスに分類することができる。

【０１１０】

検知装置１００は、ドリフトを検知した場合に、新たな訓練データセットによって、運用モデルの機械学習を再度実行する。これによって、ドリフトに対応した運用モデル５０を再度生成することができる。

【０１１１】

続いて、同一の訓練データセットを複数種類の機械学習モデルにそれぞれ入力した場合の決定境界の性質について説明する。図１８は、各機械学習モデルの決定境界の性質を示す図（１）である。図１８に示す例では、ある訓練データセット１５を用いて、サポートベクターマシン（Soft-Margin SVM）、ランダムフォレスト（Ramdom Forest）、ＮＮに対してそれぞれ機械学習を実行した例を示す。

【０１１２】

そうすると、訓練したサポートベクターマシンにデータセットを入力した場合の分布は、分布２０Ａとなり、各データは、決定境界２１Ａで第１クラス、第２クラスに分類される。訓練したランダムフォレストにデータセットを入力した場合の分布は、分布２０Ｂとなり、各データは、決定境界２１Ｂで第１クラス、第２クラスに分類される。訓練したＮＮにデータセットを入力した場合の分布は、分布２０Ｃとなり、各データは、決定境界２１Ｃで第１クラス、第２クラスに分類される。

【０１１３】

図１８に示すように、同一の訓練データセットで訓練を行った場合でも、機械学習モデルの種類によっては、決定境界の性質が違うことがわかる。

【０１１４】

図１９は、各機械学習モデルの決定境界の性質を示す図（２）である。図１９では、ある訓練データセット３５を用いて、複数種類の機械学習モデルを訓練した例を示す。ここでは、機械学習モデルとして、Nearest Neighbors、RBF SVM、Gaussian Process、Random Forest、Neural Net、Gradient Booting Tree、Naive Bayesを示す。

【０１１５】

訓練したNearest Neighborsにデータセットを入力した場合の分布は、分布４０Ａと

なる。各データは、決定境界４１Ａで第１クラス、第２クラスに分類される。訓練したNearest Neighborsのインスペクターモデルの分布は、分布４２Ａとなり、各データは、決定境界４３Ａで第１クラス、第２クラスに分類される。Ｇ４２Ａは、訓練したNearest Neighborsのインスペクターモデルを基に計算した決定境界からの距離を示す。Ｇ４２Ａでは、同じ色の等高線が、同じ距離を示す。インスペクターモデルをＮＮとする。

【０１１６】

訓練したRBF SVMにデータセットを入力した場合の分布は、分布４０Ｂとなる。各データは、決定境界４１Ｂで第１クラス、第２クラスに分類される。訓練したRBF SVMのインスペクターモデルの分布は、分布４２Ｂとなり、各データは、決定境界４３Ｂで第１クラス、第２クラスに分類される。Ｇ４２Ｂは、訓練したRBF SVMのインスペクターモデルを基に計算した決定境界からの距離を示す。インスペクターモデルをＮＮとする。

【０１１７】

訓練したGaussian Processにデータセットを入力した場合の分布は、分布４０Ｃとなる。各データは、決定境界４１Ｃで第１クラス、第２クラスに分類される。訓練したGaussian Processのインスペクターモデルの分布は、分布４２Ｃとなり、各データは、決定境界４３Ｃで第１クラス、第２クラスに分類される。Ｇ４２Ｃは、訓練したGaussian Processのインスペクターモデルを基に計算した決定境界からの距離を示す。インスペクターモデルをＮＮとする。

【０１１８】

訓練したRandom Forestにデータセットを入力した場合の分布は、分布４０Ｄとなる。各データは、決定境界４１Ｄで第１クラス、第２クラスに分類される。訓練したRandom Forestのインスペクターモデルの分布は、分布４２Ｄとなり、各データは、決定境界４３Ｄで第１クラス、第２クラスに分類される。Ｇ４２Ｄは、訓練したRandom Forestのインスペクターモデルを基に計算した決定境界からの距離を示す。インスペクターモデルをＮＮとする。

【０１１９】

訓練したNeural Netにデータセットを入力した場合の分布は、分布４０Ｅとなる。各データは、決定境界４１Ｅで第１クラス、第２クラスに分類される。訓練したNeural Netのインスペクターモデルの分布は、分布４２Ｅとなり、各データは、決定境界４３Ｅで第１クラス、第２クラスに分類される。Ｇ４２Ｅは、訓練したNeural Netのインスペクターモデルを基に計算した決定境界からの距離を示す。インスペクターモデルをＮＮとする。

【０１２０】

訓練したGradient Booting Treeにデータセットを入力した場合の分布は、分布４０Ｆとなる。各データは、決定境界４１Ｆで第１クラス、第２クラスに分類される。訓練したGradient Booting Treeのインスペクターモデルの分布は、分布４２Ｆとなり、各データは、決定境界４３Ｆで第１クラス、第２クラスに分類される。Ｇ４２Ｆは、訓練したGradient Booting Treeのインスペクターモデルを基に計算した決定境界からの距離を示す。インスペクターモデルをＮＮとする。

【０１２１】

訓練したNaive Bayesにデータセットを入力した場合の分布は、分布４０Ｇとなる。各データは、決定境界４１Ｇで第１クラス、第２クラスに分類される。訓練したNaive Bayesのインスペクターモデルの分布は、分布４２Ｇとなり、各データは、決定境界４３Ｇで第１クラス、第２クラスに分類される。Ｇ４２Ｇは、訓練したNaive Bayesのインスペクターモデルを基に計算した決定境界からの距離を示す。インスペクターモデルをＮＮとする。

【０１２２】

本実施形態で説明した検知装置１００では、機械学習モデルのアーキテクチャにかかわらず、インスペクターモデルを用いて、決定境界からの距離を近似計算することができる。

【０１２３】

次に、本実施形態に示した検知装置１００と同様の機能を実現するコンピュータのハー

ドウェア構成の一例について説明する。図２０は、本実施形態に係る検知装置と同様の機能を実現するコンピュータのハードウェア構成の一例を示す図である。

【０１２４】

図２０に示すように、コンピュータ２００は、各種演算処理を実行するＣＰＵ２０１と、ユーザからのデータの入力を受け付ける入力装置２０２と、ディスプレイ２０３とを有する。また、コンピュータ２００は、記憶媒体からプログラム等を読み取る読み取り装置２０４と、有線または無線ネットワークを介して、外部装置等との間でデータの授受を行うインタフェース装置２０５とを有する。コンピュータ２００は、各種情報を一時記憶するＲＡＭ２０６と、ハードディスク装置２０７とを有する。そして、各装置２０１～２０７は、バス２０８に接続される。

【０１２５】

ハードディスク装置２０７は、生成プログラム２０７ａ、算出プログラム２０７ｂ、取得プログラム２０７ｃ、検知プログラム２０７ｄ、分類プログラム２０７ｅを有する。ＣＰＵ２０１は、生成プログラム２０７ａ、算出プログラム２０７ｂ、取得プログラム２０７ｃ、検知プログラム２０７ｄ、分類プログラム２０７ｅを読み出してＲＡＭ２０６に展開する。

【０１２６】

生成プログラム２０７ａは、生成プロセス２０６ａとして機能する。算出プログラム２０７ｂは、算出プロセス２０６ｂとして機能する。取得プログラム２０７ｃは、取得プロセス２０６ｃとして機能する。検知プログラム２０７ｄは、検知プロセス２０６ｄとして機能する。分類プログラム２０７ｅは、分類プロセス２０６ｅとして機能する。

【０１２７】

生成プロセス２０６ａの処理は、生成部１５１の処理に対応する。算出プロセス２０６ｂの処理は、算出部１５２の処理に対応する。取得プロセス２０６ｃの処理は、取得部１５３の処理に対応する。検知プロセス２０６ｄの処理は、検知部１５４の処理に対応する。分類プロセス２０６ｄの処理は、分類部１５５の処理に対応する。

【０１２８】

なお、各プログラム２０７ａ～２０７ｅついては、必ずしも最初からハードディスク装置５０７に記憶させておかなくてもよい。例えば、コンピュータ２００に挿入されるフレキシブルディスク（ＦＤ）、ＣＤ－ＲＯＭ、ＤＶＤディスク、光磁気ディスク、ＩＣカードなどの「可搬用の物理媒体」に各プログラムを記憶させておく。そして、コンピュータ２００が各プログラム２０７ａ～２０７ｅを読み出して実行するようにしてもよい。

【符号の説明】

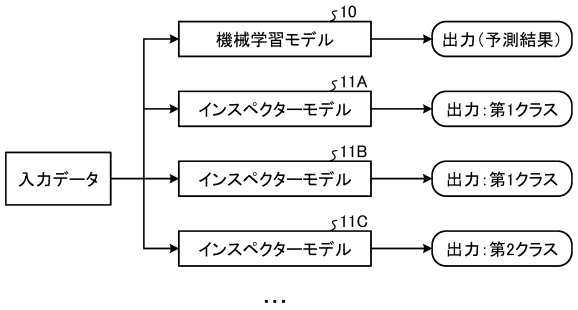
【０１２９】

１００　検知装置
１１０　通信部
１２０　入力部
１３０　出力部
１４０　記憶部
１４１　訓練データセット
１４２　運用モデルデータ
１４３　疑似サンプルテーブル
１４４　蒸留データテーブル
１４５　インスペクターモデルデータ
１４６　運用データセットテーブル
１５０　制御部
１５１　生成部
１５２　算出部
１５３　取得部
１５４　検知部

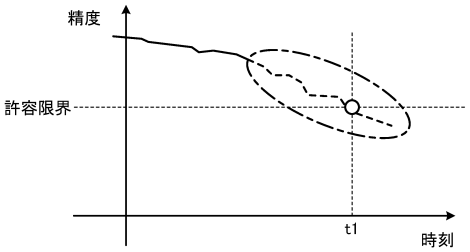
１５５　分類部

【図面】

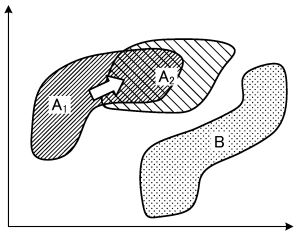
【図１】



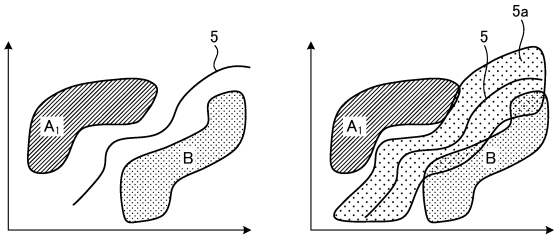
【図２】



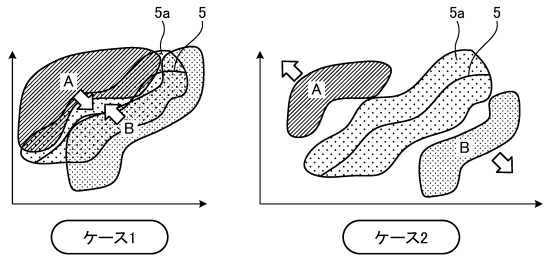
【図３】



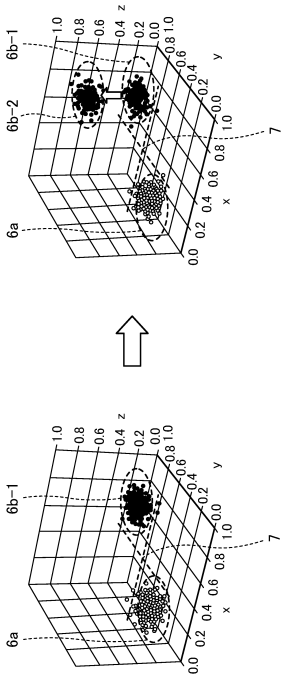
【図４】



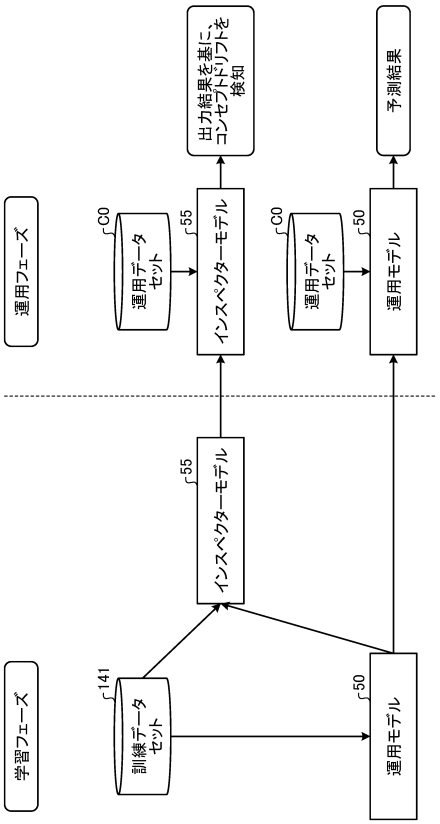
【図５】


5a
5
A
B
ケース1
5a
5
A
B
ケース2

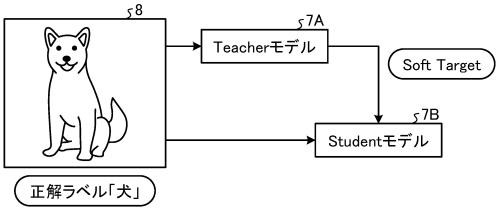
【図６】


6b−1
6b−2
6a
x
y
z

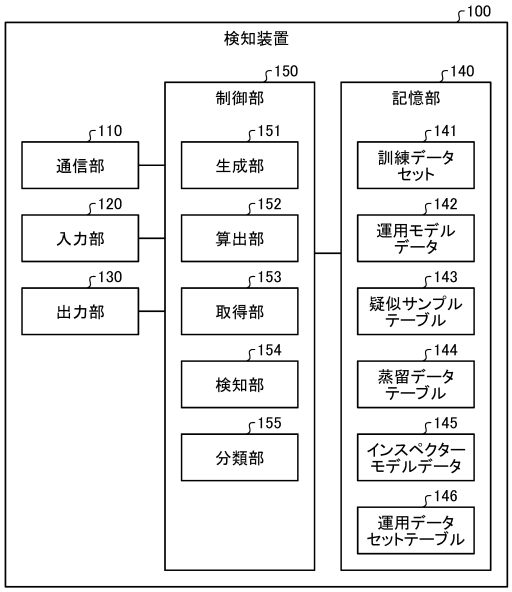
【図７】


運用フェーズ
学習フェーズ
出力結果を基に、コンセプトドリフトを検知
予測結果
運用データセット
インスペクターモデル
運用モデル
訓練データセット
C0
55
50
141

【図８】


8
7A
Teacherモデル
Soft Target
7B
Studentモデル
正解ラベル「犬」

【図９】

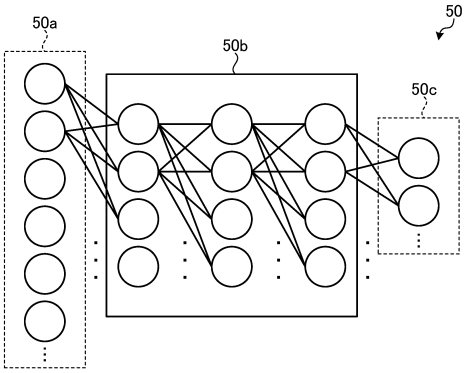


【図１０】

141

レコード番号	訓練データ	正解ラベル
1001	レコード番号「1001」訓練データ	第1クラス
1002	レコード番号「1002」訓練データ	第1クラス
1003	レコード番号「1003」訓練データ	第1クラス
・・・	・・・	・・・
1050	レコード番号「1050」訓練データ	第2クラス
1051	レコード番号「1051」訓練データ	第2クラス
・・・	・・・	・・・

【図１１】



【図１２】

143

サンプル番号	疑似サンプル
pe1001	サンプル番号「pe1001」の疑似サンプル
pe1002	サンプル番号「pe1002」の疑似サンプル
pe1003	サンプル番号「pe1003」の疑似サンプル
…	…
pe1050	サンプル番号「pe1050」の疑似サンプル
pe1051	サンプル番号「pe1051」の疑似サンプル
…	…

【図１３】

144

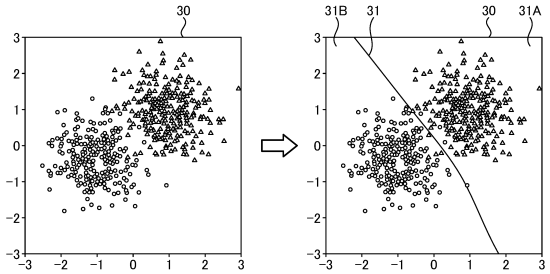
サンプル番号	疑似サンプル	ソフトターゲット
pe1001	サンプル番号「pe1001」疑似サンプル	サンプル番号「pe1001」疑似サンプルを運用モデルに入力した際の出力結果
pe1002	サンプル番号「pe1002」疑似サンプル	サンプル番号「pe1002」疑似サンプルを運用モデルに入力した際の出力結果
pe1003	サンプル番号「pe1003」疑似サンプル	サンプル番号「pe1003」疑似サンプルを運用モデルに入力した際の出力結果
…	…	…
pe1050	サンプル番号「pe1050」疑似サンプル	サンプル番号「pe1050」疑似サンプルを運用モデルに入力した際の出力結果
pe1051	サンプル番号「pe1051」疑似サンプル	サンプル番号「pe1051」疑似サンプルを運用モデルに入力した際の出力結果
…	…	…

【図１４】

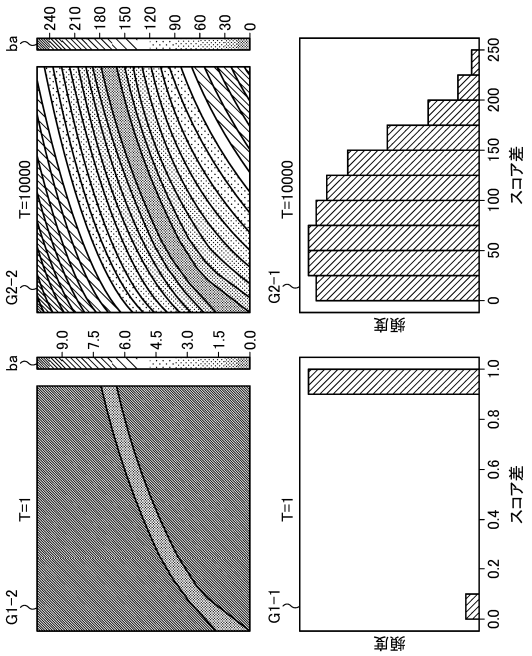
146

データ識別情報	運用データセット
C0	データ識別情報「C0」の運用データセット
C1	データ識別情報「C1」の運用データセット
C2	データ識別情報「C2」の運用データセット
C3	データ識別情報「C3」の運用データセット

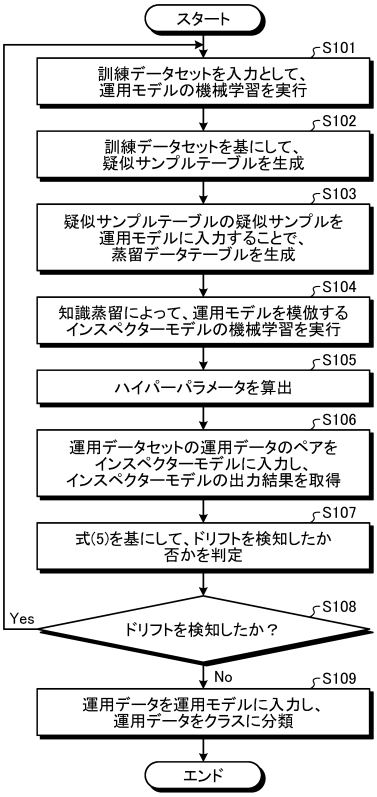
【図１５】



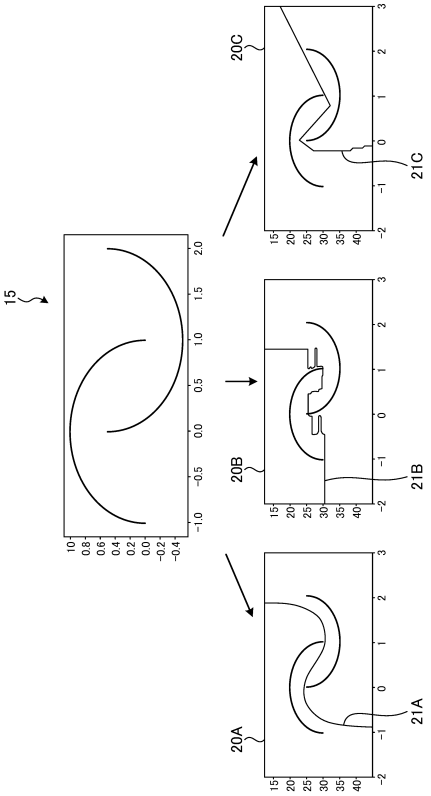
【図１６】



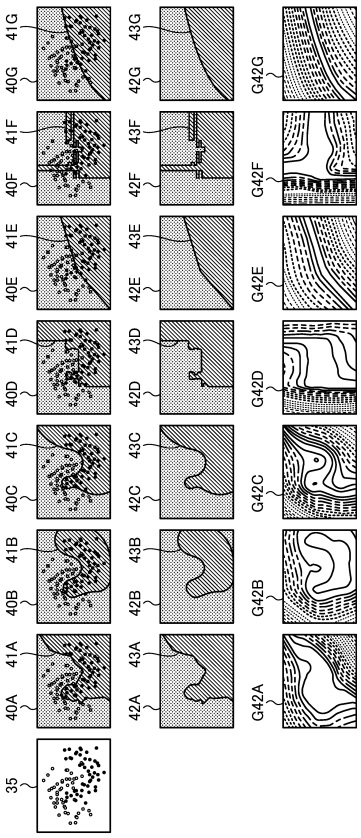
【図１７】



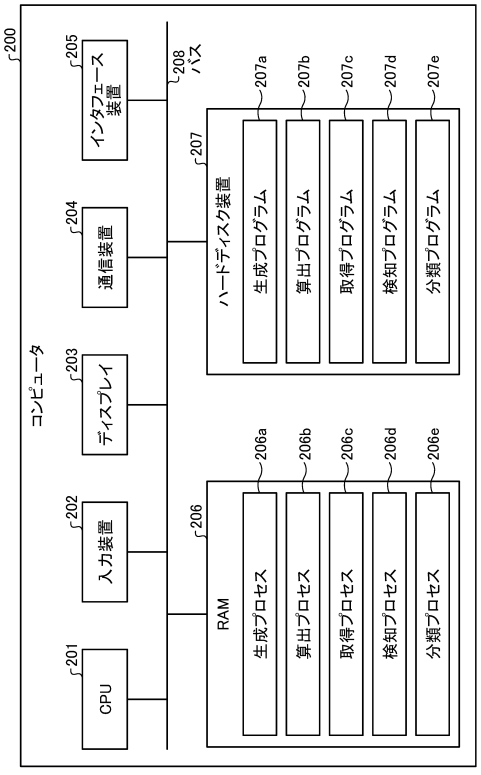
【図１８】



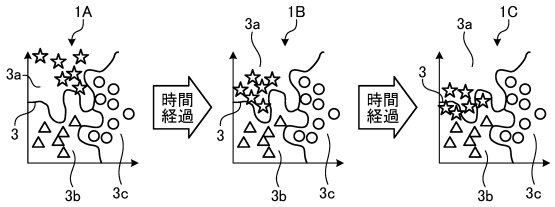
【図１９】



【図２０】



【図２１】



フロントページの続き

(56)参考文献 米国特許出願公開第２０１７／０３３０１０９（ＵＳ，Ａ１）
CERQUEIRA, Vitor et al.，"Unsupervised Concept Drift Detection Using a Student-Teacher Approach"，Lecture Notes in Computer Science [online]，スイス，Springer Nature，2020年10月15日，volume 12323，pp. 190-204，[検索日 2020.11.27], インターネット: 〈URL: https://link.springer.com/chapter/10.1007/978-3-030-61527-7_13〉
LU, Jie et al.，"Learning under Concept Drift: A Review"，IEEE Transactions on Knowledge and Data Engineering [online]，米国，IEEE，2019年12月，vol. 31, no. 12，pp. 2346-2363，[検索日 2020.11.27], インターネット: 〈URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8496795〉
石田 勉 ほか，「ラベルなし運用データに対するコンセプトドリフト検知手法の評価」，人工知能学会 第34回全国大会(2020) 論文集 [online]，一般社団法人 人工知能学会，2020年06月，セッションID: 4Rin1-05, pp. 1-4，[検索日 2020.11.27], インターネット: 〈URL: https://www.jstage.jst.go.jp/article/pjsai/JSAI2020/0/JSAI2020_4Rin105/_pdf/-char/ja〉
石井 遊哉 ほか，「Knowledge Distillationにおける温度パラメータの適正化に関する検討」，情報処理学会研究報告 コンピュータビジョンとイメージメディア（ＣＶＩＭ） [online]，情報処理学会，2019年02月28日，vol. 2019-CVIM-216, no. 10，pp. 1-5，[検索日 2019.03.01], インターネット: 〈URL: https://ipsj.ixsq.nii.ac.jp/ej/?action=repository_uri&item_id=194906&file_id=1&file_no=1〉

(58)調査した分野 (Int.Cl.，ＤＢ名)

Ｇ０６Ｎ　３／００－９９／００
Ｇ０６Ｆ　１６／００－１６／９５８
Ｇ０６Ｆ　１８／００－１８／４０
Ｇ０６Ｆ　２１／００－２１／８８
Ｇ０６Ｔ　７／００
Ｇ０６Ｖ　１０／７０－１０／８６