(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7306468号
(P7306468)

(45)発行日　令和5年7月11日(2023.7.11)

(24)登録日　令和5年7月3日(2023.7.3)

(51)国際特許分類
G06N 20/00 (2019.01)

FI
G06N 20/00

請求項の数　5　（全36頁）

(21)出願番号　特願2021-553228(P2021-553228)
(86)(22)出願日　令和1年10月24日(2019.10.24)
(86)国際出願番号　PCT/JP2019/041689
(87)国際公開番号　WO2021/079458
(87)国際公開日　令和3年4月29日(2021.4.29)
審査請求日　令和4年3月25日(2022.3.25)

(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　渡辺 順哉

最終頁に続く

(54)【発明の名称】　検出方法、検出プログラムおよび情報処理装置

(57)【特許請求の範囲】

【請求項１】

コンピュータが実行する検出方法であって、

第１クラスおよび第２クラスに対応する複数の訓練データを用いて、監視対象となる運用モデルを訓練し、

前記運用モデルの知識蒸留を基にして、前記第１クラスの領域と前記第２クラスの領域との決定境界から運用データまでの距離を算出するインスペクターモデルを訓練することで、前記インスペクターモデルに、前記決定境界を学習させ、

前記複数の訓練データおよび複数の運用データを前記インスペクターモデルに入力した結果を基にして、データの傾向の時間変化に起因する前記運用モデルの出力結果の変化を検出する

処理を実行することを特徴とする検出方法。

【請求項２】

前記変化を検出する処理は、前記複数の訓練データを前記インスペクターモデルに入力した結果を基にして、前記複数の訓練データのうち、前記決定境界から任意に設定された範囲内に含まれる訓練データの第一割合を算出し、

前記複数の運用データを前記インスペクターモデルに入力した結果を基にして、前記複数の運用データのうち、前記決定境界から任意に設定された範囲内に含まれる運用データの第二割合を算出し、

前記第一割合と前記第二割合とを基にして、前記運用モデルの出力結果の変化を検出す

ることを特徴とする請求項１に記載の検出方法。
【請求項３】
前記運用モデルにデータを入力して、入力したデータが、前記第１クラスに対応するのか、前記第２クラスに対応するのかを判定し、判定結果を入力したデータに対応付ける処理を、複数のデータついて実行することで、訓練データセットを生成する処理を更に実行し、
前記インスペクターモデルを作成する処理は、前記訓練データセットを用いて、前記決定境界を学習することを特徴とする請求項２に記載の検出方法。
【請求項４】
コンピュータに、
第１クラスおよび第２クラスに対応する複数の訓練データを用いて、監視対象となる運用モデルを訓練し、
前記運用モデルの知識蒸留を基にして、前記第１クラスの領域と前記第２クラスの領域との決定境界から運用データまでの距離を算出するインスペクターモデルを訓練することで、前記インスペクターモデルに、前記決定境界を学習させ、
前記複数の訓練データおよび複数の運用データを前記インスペクターモデルに入力した結果を基にして、データの傾向の時間変化に起因する前記運用モデルの出力結果の変化を検出する
処理を実行させることを特徴とする検出プログラム。
【請求項５】
第１クラスおよび第２クラスに対応する複数の訓練データを用いて、監視対象となる運用モデルを訓練する学習部と、
前記運用モデルの知識蒸留を基にして、前記第１クラスの領域と前記第２クラスの領域との決定境界から運用データまでの距離を算出するインスペクターモデルを訓練することで、前記インスペクターモデルに、前記決定境界を学習させる作成部と、
前記複数の訓練データおよび複数の運用データを前記インスペクターモデルに入力した結果を基にして、データの傾向の時間変化に起因する前記運用モデルの出力結果の変化を検出する検出部と
を有することを特徴とする情報処理装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、検出方法等に関する。
【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教師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３２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、教師データに基づき、予め学習されているものとする。教師データには、訓練データと、検証データとが含まれる。
【０００４】
図３２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、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出する技術として、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、画像データ等の高次元データに対して、T^2統計量を適用することが難しく、機械学習モデルの精度劣化を検知することができない。
【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、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、異常データを上手く検知することができず、機械学習モデルの精度劣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検出することができる検出方

法、検出プログラムおよび情報処理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】
第１の案では、コンピュータが次の処理を実行する。コンピュータは、第１クラスまたは第２クラスに対応する複数の訓練データを用いて、監視対象となる運用モデルを学習する。コンピュータは、運用モデルの知識蒸留を基にして、第１クラスの領域と第２クラスの領域との決定境界を学習すると共に、決定境界から運用データまでの距離を算出するインスペクターモデルを作成する。コンピュータは、複数の訓練データおよび複数の運用データをインスペクターモデルに入力した結果を基にして、データの傾向の時間変化に起因する運用モデルの出力結果の変化を検出する。
【発明の効果】
【００１６】
機械学習モデルの精度劣化を検出することができる。
【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、知識蒸留を説明するための図である。
【図６】図６は、決定境界周辺の危険領域の算出手法を説明するための図である。
【図７】図７は、各機械学習モデルの決定境界の性質を示す図である。
【図８】図８は、各インスペクターモデルの決定境界の可視化結果を示す図である。
【図９】図９は、各インスペクターモデルによる危険領域を可視化した図である。
【図１０】図１０は、本実施例１に係る情報処理装置の構成を示す機能ブロック図である。
【図１１】図１１は、本実施例１に係る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１２】図１２は、本実施例１に係る機械学習モデルの一例を説明するための図である。
【図１３】図１３は、本実施例１に係る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、本実施例１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、作成部の処理を説明するための図（１）である。
【図１７】図１７は、作成部の処理を説明するための図（２）である。
【図１８】図１８は、本実施例１に係る検出部の処理を説明するための図（１）である。
【図１９】図１９は、本実施例１に係る検出部の処理を説明するための図（２）である。
【図２０】図２０は、本実施例１に係る情報処理装置の処理手順を示すフローチャートである。
【図２１】図２１は、本実施例２に係る情報処理装置の処理を説明するための図である。
【図２２】図２２は、本実施例２に係る情報処理装置の構成を示す機能ブロック図である。
【図２３】図２３は、本実施例２に係る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２４】図２４は、本実施例２に係る機械学習モデルの一例を説明するための図である。
【図２５】図２５は、本実施例２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２６】図２６は、インスペクターモデルの決定境界および危険領域の一例を示す図である。
【図２７】図２７は、本実施例２に係る情報処理装置の処理手順を示すフローチャートである。
【図２８】図２８は、本実施例３に係る情報処理装置の処理を説明するための図である。
【図２９】図２９は、本実施例３に係る情報処理装置の構成を示す機能ブロック図である。

【図３０】図３０は、本実施例３に係る情報処理装置の処理手順を示すフローチャートである。
【図３１】図３１は、本実施例に係る情報処理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３２】図３２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、本願の開示する検出方法、検出プログラムおよび情報処理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。
【実施例１】
【００１９】
本実施例１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、教師データを用いて機械学習した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、教師データには、訓練データと、検証データとが含まれる。訓練データは、機械学習モデル１０のパラメータを機械学習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。検証データは、機械学習モデル１０を検証する場合に用いられるデータである。
【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を作成している。
【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、再学習が必要となる。

【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を作成している。そして、複数のインスペクターモデル１１Ａ～１１Ｃを作成するためには、機械学習モデル１０や、機械学習モデル１０の学習時に用いた、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が特定の学習モデルであることが求められる。

【００２９】

そうすると、機械学習モデル１０の精度劣化を検知する手法が、機械学習モデルに依存してしまう。機械学習モデルの分類アルゴリズムには、ＮＮ（Neural Network）、決定木、ｋ近傍法、サポートベクターマシン等様々な分類アルゴリズムが該当するため、分類アルゴリズム毎に、どの検知手法が精度劣化の検知に適する手法であるかを試行錯誤する必要がある。

【００３０】

すなわち、どのような分類アルゴリズムであっても、汎用的に使用可能なインスペクターモデルを作成し、機械学習モデル１０の精度劣化を検知することが望ましい。

【００３１】

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を学習することで、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（または減少）したか否かを特定し、運用データの数が増加（または減少）した場合に、精度劣化を検出する。

【００３２】

以下の説明において、訓練データは、監視対象となる機械学習モデルを学習する場合に用いるデータである。運用データは、機械学習モデルを用いて、各分類クラスに分類するデータであり、運用開始時からの時間経過に応じて特徴量が変化するものとする。

【００３３】

本実施例１に係る情報処理装置は、知識蒸留（ＫＤ：Knowledge Distiller）を用いて、決定境界５の危険領域５ａに含まれる運用データの数の増減を算出し、機械学習モデルの精度劣化を検出する。

【００３４】

図５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。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ＡおよびStudentモデル７ＢをＮＮとするが、これに限定されるものではない。

【００３５】

情報処理装置は、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のパラメータを学習（誤差逆伝播法に

よる学習）する。また、情報処理装置は、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のパラメータを学習する。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。

【００３６】

上記のように、Teacherモデル７Ａに関する学習を、訓練データ６とハードターゲットとを用いて学習し、Studentモデル７Ｂに関する学習を、訓練データ６とソフトターゲットとを用いて学習する手法を、知識蒸留と呼ぶ。情報処理装置は、他の訓練データについても同様にして、TeacherモデルＡおよびStudentモデル７Ｂを学習する。

【００３７】

ここで、データ空間を入力としたソフトターゲットで、Studentモデル７Ｂの学習を考える。Teacherモデル７Ａと、Studentモデル７Ｂとを異なるモデルで構築すれば、Studentモデル７Ｂの出力結果は、Teacherモデル７Ａの出力結果の決定境界に類似するように学習される。そうすると、Teacherモデル７Ａを監視対象の機械学習モデル、Studentモデル７Ｂをインスペクターモデルとして扱うことが可能となる。Teacherモデル７Ａのモデルアーキテクチャを絞らないことで、汎用的に使用可能なインスペクターモデルを作成することができる。

【００３８】

図６は、決定境界周辺の危険領域の算出手法を説明するための図である。本実施例１に係る情報処理装置は、特徴量空間の決定境界５が直線になるような高次元空間（再生核ヒルベルト空間）Ｈｋにデータ（ソフトターゲット）を射影して、危険領域５ａを算出する。たとえば、データ８を入力した場合に、高次元空間Ｈｋの決定境界５と、データ８との距離（符号付きの距離）m_8を算出するインスペクターモデルを構築する。危険領域５ａの幅を幅ｍとし、距離m_8がｍ未満である場合には、データ８は、危険領域５ａに含まれることを意味する。距離（ノルム）の計算は、再生核ヒルベルト空間の内積によって計算され、カーネルトリックに対応する。距離（ノルム）は、式（１）によって定義される。

【００３９】

【数１】

$$\|f\| = \sqrt{\langle f, f \rangle} \quad \cdots (1)$$

【００４０】

情報処理装置は、インスペクターモデルを、Hard-Margin RBF（Radial Basis Function）カーネルSVM（Support Vector Machine）によって構築する。情報処理装置は、再生核ヒルベルト空間に、決定境界５が直線になるようにデータ空間を射影する。危険領域５ａの幅ｍは、精度劣化に関する検知の感度であり、決定境界５付近のデータ密度で決定される。

【００４１】

たとえば、情報処理装置は、ソフトターゲットの領域を領域Ｘおよび領域Ｙに分類する。情報処理装置は、領域Ｘおよび領域Ｙを、再生核ヒルベルト空間に射影し、決定境界５側に一番近いサポートベクトルＸａ、Ｙａを特定する。情報処理装置は、サポートベクトルＸａおよび決定境界５のマージンと、サポートベクトルＹａおよび決定境界５のマージンとの差が最小となるように、決定境界５を特定する。つまり、情報処理装置は、監視した機械学習モデルの決定境界５との乖離を損失として学習しながら、ユークリッド空間上の決定境界付近の空間をねじ曲げることに相当する処理を実行する。

【００４２】

ここで、本実施例１に係る情報処理装置が、上記処理によって作成したインスペクターモデルを用いて、監視対象の機械学習モデルの精度劣化を検知する処理の一例について説

明する。なお、機械学習モデルは、複数の訓練データによって、学習済みとする。以下の説明では、複数の訓練データを「訓練データセット」と表記する。

【００４３】

情報処理装置は、訓練データセットに含まれる各訓練データを、インスペクターモデルに入力し、全訓練データのうち、危険領域５ａに含まれる訓練データの割合を算出しておく。以下の説明において、全訓練データのうち、危険領域５ａに含まれる訓練データの割合を「第一割合」と表記する。

【００４４】

情報処理装置は、機械学習モデルの運用開始時から時間経過した後に、運用データセットを取得する。運用データセットには、複数の運用データが含まれる。情報処理装置は、運用データセットに含まれる各運用データを、インスペクターモデルに入力し、全運用データのうち、危険領域５ａに含まれる運用データの割合を算出する。以下の説明において、全運用データのうち、危険領域５ａに含まれる訓練データの割合を「第二割合」と表記する。

【００４５】

情報処理装置は、第一割合と第二割合とを比較して、第二割合が増加または減少した場合、機械学習モデルの精度劣化を検知する。第一割合を基準として、第二割合が変化したということは、運用開始時と比較して、多くの運用データが、危険領域５ａに含まれており、コンセプトドリフトが発生していることを示す。情報処理装置は、時間経過に伴って、運用データセットを取得し、上記処理を繰り返し実行する。これによって、どのような分類アルゴリズムであっても、汎用的に使用可能なインスペクターモデルを作成し、機械学習モデルの精度劣化を検知することができる。

【００４６】

次に、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７は、各機械学習モデルの決定境界の性質を示す図である。図７に示す例では、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をそれぞれ学習する。

【００４７】

そうすると、学習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。学習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。学習した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。

【００４８】

図７に示すように、同一の訓練データセット１５で学習を行った場合でも、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。

【００４９】

続いて、各機械学習モデルを用いた知識蒸留によって、インスペクターモデルを作成した場合の決定境界の一例について説明する。説明の便宜上、機械学習モデル（サポートベクターマシン）を用いた知識蒸留によって作成したインスペクターモデルを、第１インスペクターモデルと表記する。機械学習モデル（ランダムフォレスト）を用いた知識蒸留によって作成したインスペクターモデルを、第２インスペクターモデルと表記する。機械学習モデル（ＮＮ）を用いた知識蒸留によって作成したインスペクターモデルを、第３インスペクターモデルと表記する。

【００５０】

図８は、各インスペクターモデルの決定境界を可視化した結果を示す図である。情報処理装置は、分布２０Ａを基にして、第１インスペクターモデルを作成すると、第１インスペクターモデルの分布は、２２Ａに示すものとなり、決定境界は、決定境界２３Ａとなる。

【００５１】
情報処理装置は、分布２０Ｂを基にして、第２インスペクターモデルを作成すると、第２インスペクターモデルの分布は、２２Ｂに示すものとなり、決定境界は、決定境界２３Ｂとなる。情報処理装置は、分布２０Ｃを基にして、第３インスペクターモデルを作成すると、第３インスペクターモデルの分布は、２２Ｃに示すものとなり、決定境界は、決定境界２３Ｃとなる。
【００５２】
図９は、各インスペクターモデルによる危険領域を可視化した図である。第１インスペクターモデルの決定境界２３Ａを基にした危険領域は、危険領域２４Ａとなる。第２インスペクターモデルの決定境界２３Ｂを基にした危険領域は、危険領域２４Ｂとなる。第３インスペクターモデルの決定境界２３Ｃを基にした危険領域は、危険領域２４Ｃとなる。
【００５３】
次に、本実施例１に係る情報処理装置の構成について説明する。図１０は、本実施例１に係る情報処理装置の構成を示す機能ブロック図である。図１０に示すように、情報処理装置１００は、通信部１１０と、入力部１２０と、表示部１３０と、記憶部１４０と、制御部１５０とを有する。
【００５４】
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処理部である。通信部１１０は、通信装置の一例である。後述する制御部１５０は、通信部１１０を介して、外部装置とデータをやり取りする。
【００５５】
入力部１２０は、情報処理装置１００に対して各種の情報を入力するための入力装置である。入力部１２０は、キーボードやマウス、タッチパネル等に対応する。
【００５６】
表示部１３０は、制御部１５０から出力される情報を表示する表示装置である。表示部１３０は、液晶ディスプレイ、有機ＥＬ（Electro Luminescence）ディスプレイ、タッチパネル等に対応する。
【００５７】
記憶部１４０は、教師データ１４１、機械学習モデルデータ１４２、蒸留データテーブル１４３、インスペクターモデルデータ１４４、運用データテーブル１４５を有する。記憶部１４０は、ＲＡＭ（Random Access Memory）、フラッシュメモリ（Flash Memory）などの半導体メモリ素子や、ＨＤＤ（Hard Disk Drive）などの記憶装置に対応する。
【００５８】
教師データ１４１は、訓練データセット１４１ａと、検証データ１４１ｂを有する。訓練データセット１４１ａは、訓練データに関する各種の情報を保持する。
【００５９】
図１１は、本実施例１に係る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する。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。
【００６０】
検証データ１４１ｂは、訓練データセット１４１ａによって学習された機械学習モデルを検証するためのデータである。検証データ１４１ｂは、正解ラベルが付与される。たとえば、検証データ１４１ｂを、機械学習モデルに入力した場合に、機械学習モデルから出力される出力結果が、検証データ１４１ｂに付与される正解ラベルに一致する場合、訓練データセット１４１ａによって、機械学習モデルが適切に学習されたことを意味する。
【００６１】

機械学習モデルデータ１４２は、機械学習モデルのデータである。本実施例１に機械学習モデルは、所定の分類アルゴリズムによって、入力データを、第１クラスまたは第２クラスに分類する機械学習モデルである。分類アルゴリズムは、ＮＮ、ランダムフォレスト、ｋ近傍法、サポートベクターマシン等のうち、いずれの分類アルゴリズムであってもよい。

【００６２】

ここでは一例として、機械学習モデルを、ＮＮとして説明を行う。図１２は、機械学習モデルの一例を説明するための図である。図１２に示すように、機械学習モデル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、エッジは、重みを持つ。以下の説明では、バイアス値、重みを「パラメータ」と表記する。

【００６３】

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のノード５１ａ，５１ｂから、各クラスの確率が出力される。たとえば、ノード５１ａから、第１クラスの確率が出力される。ノード５１ｂから、第２クラスの確率が出力される。

【００６４】

蒸留データテーブル１４３は、データセットの各データを、機械学習モデル５０に入力した場合の出力結果（ソフトターゲット）を格納するテーブルである。図１３は、本実施例１に係る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１４３は、レコード番号と、入力データと、ソフトターゲットとを対応付ける。レコード番号は、入力データと、ソフトターゲットとの組を識別する番号である。入力データは、学習された機械学習モデル５０の決定境界（決定境界を含む特徴空間）を基にして、作成部１５２に選択されるデータである。

【００６５】

ソフトターゲットは、入力データを学習済みの機械学習モデル５０に入力した場合に出力されるものである。たとえば、本実施例１に係るソフトターゲットは、第１クラスまたは第２クラスのうち、いずれかの分類クラスを示すものとする。

【００６６】

インスペクターモデルデータ１４４は、Hard-Margin RBFカーネルSVMによって構築されたインスペクターモデルのデータである。以下の説明では、Hard-Margin RBFカーネルSVMを「ｋＳＶＭ」と表記する。かかるインスペクターモデルに、データを入力すると、符号付きの距離の値が出力される。たとえば、符号がプラスであれば、入力したデータは第１クラスに分類される。符号がマイナスであれば、データは、第２クラスに分類される。距離は、データと決定境界との距離を示す。

【００６７】

運用データテーブル１４５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、運用データテーブル１４５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、複数の運用データが含まれる。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。

【００６８】

図１０の説明に戻る。制御部１５０は、学習部１５１と、作成部１５２と、検出部１５３と、予測部１５４とを有する。制御部１５０は、ＣＰＵ（Central Processing Unit）やＭＰＵ（Micro Processing Unit）などによって実現できる。また、制御部１５０は、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）などのハードワイヤードロジックによっても実現できる。

【００６９】
学習部１５１は、訓練データセット１４１ａを取得し、訓練データセット１４１ａを基にして、機械学習モデル５０のパラメータを学習する処理部である。たとえば、学習部１５１は、訓練データセット１４１ａの訓練データを、機械学習モデル５０の入力層に入力した場合、出力層の各ノードの出力結果が、入力した訓練データの正解ラベルに近づくように、機械学習モデル５０のパラメータを更新する（誤差逆伝播法による学習）。学習部１５１は、訓練データセット１４１ａに含まれる各訓練データについて、上記処理を繰り返し実行する。また、学習部１５１は、検証データ１４１ｂを用いて、機械学習モデル５０の検証を行ってもよい。学習部１５１は、学習済みの機械学習モデル５０のデータ（機械学習モデルデータ１４２）を、記憶部１４０に登録する。機械学習モデル５０は、「運用モデル」の一例である。

【００７０】
図１５は、本実施例１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１４１ａ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。

【００７１】
たとえば、訓練データセット１４１ａによって、機械学習モデル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、機械学習モデル５０が、ＮＮである場合、機械学習モデル５０に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。

【００７２】
作成部１５２は、機械学習モデル５０の知識蒸留を基にして、モデル適用領域３１Ａとモデル適用領域３１Ｂとの決定境界３１を学習した、インスペクターモデルを作成する処理部である。このインスペクターモデルにデータ（訓練データまたは運用データ）を入力すると、決定境界３１とデータとの距離（符号付きの距離の値）が出力される。

【００７３】
作成部１５２は、蒸留データテーブル１４３を生成する処理、インスペクターモデルデータ１４４を作成する処理を実行する。

【００７４】
作成部１５２が、蒸留データテーブル１４３を生成する処理について説明する。図１６は、作成部の処理を説明するための図（１）である。作成部１５２は、機械学習モデルデータ１４２を用いて、機械学習モデル５０を実行し、特徴空間３０上の各データを、機械学習モデル５０に入力する。これにより、特徴空間３０の各データが、第１クラスに分類されるか、第２クラスに分類するのかを特定する。かかる処理を実行することで、作成部１５２は、特徴空間をモデル適用領域３１Ａと、モデル適用領域３１Ｂとに分類し、決定境界３１を特定する。

【００７５】
作成部１５２は、特徴空間３０上において、所定間隔毎に複数の縦線と横線とを配置する。所定間隔毎に複数の縦線と横線とを配置したものを「グリッド」と表記する。グリッドの幅は、予め設定されているものとする。作成部１５２は、グリッドの交点座標のデータを選択し、選択したデータを、機械学習モデル５０に出力することで、選択したデータに対応するソフトターゲットを算出する。作成部１５２は、選択したデータ（入力データ）と、ソフトターゲットとを対応付けて、蒸留データテーブル１４３に登録する。作成部１５２は、グリッドの各交点座標のデータについても、上記処理を繰り返し実行すること

で、蒸留データテーブル１４３を生成する。

【００７６】

続いて、作成部１５２が、インスペクターモデルデータ１４４を作成する処理について説明する。図１７は、作成部の処理を説明するための図（２）である。作成部１５２は、蒸留データテーブル１４３に登録された入力データと、ソフトターゲットとの関係を基にして、ｋＳＶＭによって構築されたインスペクターモデル３５を作成する。作成部１５２は、作成したインスペクターモデル３５のデータ（インスペクターモデルデータ１４４）を、記憶部１４０に登録する。

【００７７】

たとえば、作成部１５２は、蒸留データテーブル１４３に格納された各入力データを、再生核ヒルベルト空間に射影する。作成部１５２は、再生核ヒルベルト空間に含まれる第１クラスの入力データのうち、決定境界３１に最も近い入力データを、第１サポートベクトルとして選択する。作成部１５２は、再生核ヒルベルト空間に含まれる第２クラスの入力データのうち、決定境界３１に最も近い入力データを、第２サポートベクトルとして選択する。作成部１５２は、第１サポートベクトルと、第２サポートベクトルとの中間を通る決定境界３１を特定することで、インスペクターモデル（ｋＳＶＭ）のハイパーパラメータを特定する。再生核ヒルベルト空間において、決定境界３１は直線となり、決定境界３１からの距離がｍとなる領域を、危険領域３２に設定する。距離ｍは、決定境界３１と、第１サポートベクトル（第２サポートベクトル）との距離である。

【００７８】

図１０の説明に戻る。検出部１５３は、インスペクターモデル３５を実行して、機械学習モデル５０の精度劣化を検出する処理部である。検出部１５３は、訓練データセット１４１ａの各訓練データを、インスペクターモデル３５に入力する。検出部１５３が、訓練データをインスペクターモデル３５に入力すると、特徴空間上の決定境界３１と訓練データとの距離（ノルム）が出力される。

【００７９】

検出部１５３は、決定境界３１と訓練データとの距離がｍ未満である場合、かかる訓練データが危険領域３２に含まれると判定する。検出部１５３は、訓練データセット１４１ａに含まれる各訓練データについて、上記処理を繰り返し実行する。検出部１５３は、全訓練データのうち、危険領域３２に含まれる訓練データの割合を「第一割合」として算出する。

【００８０】

検出部１５３は、運用データテーブル１４５に格納された運用データセットを選択し、運用データセットの各運用データを、インスペクターモデル３５に入力する。検出部１５３が、運用データをインスペクターモデル３５に入力すると、特徴空間上の決定境界３１と運用データとの距離（ノルム）が出力される。

【００８１】

検出部１５３は、決定境界３１と運用データとの距離がｍ未満である場合、かかる運用データが危険領域３２に含まれると判定する。検出部１５３は、運用データセットに含まれる各運用データについて、上記処理を繰り返し実行する。検出部１５３は、全運用データのうち、危険領域３２に含まれる運用データの割合を「第二割合」として算出する。

【００８２】

検出部１５３は、第一割合と、第二割合とを比較し、第一割合に対して第二割合が変化した場合に、コンセプトドリフトが発生したと判定し、機械学習モデル５０の精度劣化を検出する。たとえば、検出部１５３は、第一割合と第二割合との絶対値の差分が、閾値以上となる場合に、コンセプトドリフトが発生したと判定する。

【００８３】

図１８および図１９は、本実施例１に係る検出部の処理を説明するための図である。図１８は、第一割合の一例を示す。たとえば、検出部１５３は、訓練データセット１４１ａの各訓練データをインスペクターモデル３５に入力すると、第一割合は「０．０２」とな

る場合を示している。

【００８４】

図１９は、第二割合の一例を示す。たとえば、運用データセットＣ０の各運用データをインスペクターモデル３５に入力すると、第二割合は「０．０２」となる。第一割合と、運用データセットＣ０の第二割合とは同じであるため、運用データセットＣ０において、コンセプトドリフトは発生していない。このため、検出部１５３は、運用データセットＣ０について、機械学習モデル５０の精度劣化を検出しない。

【００８５】

たとえば、運用データセットＣ１の各運用データをインスペクターモデル３５に入力すると、第二割合は「０．０９」となる。第一割合と比較して、運用データセットＣ１の第二割合が増加しており、運用データセットＣ１において、コンセプトドリフトは発生している。このため、検出部１５３は、運用データセットＣ１について、機械学習モデル５０の精度劣化を検出する。

【００８６】

たとえば、運用データセットＣ２の各運用データをインスペクターモデル３５に入力すると、第二割合は「０．０５」となる。第一割合と比較して、運用データセットＣ２の第二割合が増加しており、運用データセットＣ２において、コンセプトドリフトは発生している。このため、検出部１５３は、運用データセットＣ２について、機械学習モデル５０の精度劣化を検出する。

【００８７】

たとえば、運用データセットＣ３の各運用データをインスペクターモデル３５に入力すると、第二割合は「０．００２５」となる。第一割合と比較して、運用データセットＣ３の第二割合が減少しており、運用データセットＣ３において、コンセプトドリフトは発生している。このため、検出部１５３は、運用データセットＣ３について、機械学習モデル５０の精度劣化を検出する。

【００８８】

検出部１５３は、機械学習モデル５０の精度劣化を検出した場合には、精度劣化を検出した旨の情報を、表示部１３０に表示してもよいし、外部装置（図示略）に、精度劣化を検出した旨を通知してもよい。検出部１５３は、精度劣化を検出した根拠となる運用データセットのデータ識別情報を、表示部１３０に出力して表示させてもよい。また、検出部１５３は、精度劣化を検出した旨を学習部１５１に通知して、機械学習モデルデータ１４２を再学習させてもよい。この場合、学習部１５１は、新たに指定される訓練データセットを用いて、機械学習モデル５０を再学習する。

【００８９】

検出部１５３は、機械学習モデル５０の精度劣化を検出しない場合には、精度劣化を検出していない旨の情報を予測部１５４に出力する。

【００９０】

予測部１５４は、機械学習モデル５０の精度劣化が検出されていない場合、機械学習モデル５０を実行して、運用データセットを入力し、各運用データの分類クラスを予測する処理部である。予測部１５４は、予測結果を、表示部１３０に出力して表示させてもよいし、外部装置に送信してもよい。

【００９１】

次に、本実施例１に係る情報処理装置１００の処理手順の一例について説明する。図２０は、本実施例１に係る情報処理装置の処理手順を示すフローチャートである。図２０に示すように、情報処理装置１００の学習部１５１は、訓練データセット１４１ａを基にして、機械学習モデル５０を学習する（ステップＳ１０１）。

【００９２】

情報処理装置１００の作成部１５２は、知識蒸留を用いて、蒸留データテーブル１４３を生成する（ステップＳ１０２）。作成部１５２は、蒸留データテーブル１４３を基にして、インスペクターモデルを生成する（ステップＳ１０３）。

【００９３】
情報処理装置１００の検出部１５３は、訓練データセット１４１ａの各訓練データをインスペクターモデルに入力し、第一割合を算出する（ステップＳ１０４）。情報処理装置１００は、運用データセットの各運用データをインスペクターモデルに入力し、第二割合を算出する（ステップＳ１０５）。
【００９４】
情報処理装置１００の検出部１５３は、第一割合と第二割合とを基にして、コンセプトドリフトが発生したか否かを判定する（ステップＳ１０６）。情報処理装置１００は、コンセプトドリフトが発生した場合には（ステップＳ１０７，Ｙｅｓ）、ステップＳ１０８に移行する。一方、情報処理装置１００は、コンセプトドリフトが発生していない場合には（ステップＳ１０７，Ｎｏ）、ステップＳ１０９に移行する。
【００９５】
ステップＳ１０８以降の処理について説明する。学習部１５１は、新たな訓練データセットによって、機械学習モデル５０を再学習し（ステップＳ１０８）、ステップＳ１０２に移行する。
【００９６】
ステップＳ１０９以降の処理について説明する。情報処理装置１００の予測部１５４は、運用データセットを、機械学習モデルに入力し、各運用データの分類クラスを予測する（ステップＳ１０９）。予測部１５４は、予測結果を出力する（ステップＳ１１０）。
【００９７】
次に、本実施例１に係る情報処理装置１００の効果について説明する。情報処理装置１００は、訓練データセット１４１ａを基にして、機械学習モデル５０を生成し、知識蒸留を用いて、インスペクターモデルを作成する。情報処理装置１００は、インスペクターモデルに訓練データセットを入力した場合の第一割合と、運用データセットを入力した場合の第二割合とを算出し、第一割合と第二割合とを基にして、機械学習モデル５０の精度劣化を検出する。これによって、機械学習モデルの精度劣化を検出することができる。
【００９８】
情報処理装置１００は、第一割合と第二割合とを比較して、第二割合が増加または減少した場合、機械学習モデルの精度劣化を検知する。第一割合を基準として、第二割合が変化したということは、運用開始時と比較して、多くの運用データが、危険領域に含まれており、コンセプトドリフトが発生していることを示す。情報処理装置１００は、時間経過に伴って、運用データセットを取得し、上記処理を繰り返し実行する。これによって、どのような分類アルゴリズムであっても、汎用的に使用可能なインスペクターモデルを作成し、機械学習モデルの精度劣化を検知することができる。
【００９９】
たとえば、本実施例１に係る情報処理装置１００は、機械学習モデル５０を用いた知識蒸留によって、インスペクターモデル（カーネルＳＶＭ）を構築するため、図７～図９で説明したように、どのような分類アルゴリズムであっても、汎用的に使用可能なインスペクターモデルを作成できる。
【実施例２】
【０１００】
本実施例２に係る情報処理装置は、３種類以上の分類クラスについて、分類クラス毎に１対他の蒸留を行うことによって、監視対象となる機械学習モデルの精度劣化を検知する。また、情報処理装置は、精度劣化を検知した場合に、どの分類クラスに影響が出ているのかを特定する。
【０１０１】
図２１は、本実施例２に係る情報処理装置の処理を説明するための図である。本実施例２では、第１クラスに対応する第１訓練データセット４０Ａと、第２クラスに対応する第２訓練データセット４０Ｂと、第３クラスに対応する第３訓練データセット４０Ｃとを用いて説明する。

【０１０２】
ここでは、第１訓練データセット４０Ａに含まれる複数の第１訓練データをバツ印で示す。第２訓練データセット４０Ｂに含まれる複数の第２訓練データを三角印で示す。第３訓練データセット４０Ｃに含まれる複数の第３訓練データを丸印で示す。
【０１０３】
情報処理装置は、知識蒸留を用いて、「第１訓練データセット４０Ａ」と、「第２訓練データセット４０Ｂおよび第２訓練データセット４０Ｂ」との決定境界４１Ａを学習したインスペクターモデルＭ１を作成する。インスペクターモデルＭ１では、決定境界４１Ａ周辺の危険領域４２Ａを設定する。
【０１０４】
情報処理装置は、知識蒸留を用いて、「第２訓練データセット４０Ｂ」と、「第１訓練データセット４０Ａおよび第３訓練データセット４０Ｃ」との決定境界４１Ｂを学習したインスペクターモデルＭ２を作成する。インスペクターモデルＭ１では、決定境界４１Ｂ周辺の危険領域４２Ｂを設定する。
【０１０５】
情報処理装置は、知識蒸留を用いて、「第３訓練データセット４０Ｃ」と、「第１訓練データセット４０Ａおよび第２訓練データセット４０Ｂ」との決定境界４１Ｃを学習したインスペクターモデルＭ３を作成する。インスペクターモデルＭ３では、決定境界４１Ｃ周辺の危険領域４２Ｃを設定する。
【０１０６】
情報処理装置は、インスペクターモデルＭ１，Ｍ２，Ｍ３それぞれについて、第一割合および第二割合をそれぞれ算出する。以下の説明において、インスペクターモデルＭ１を用いて算出した第一割合を「割合Ｍ１－１」と表記し、インスペクターモデルＭ１を用いて算出した第二割合を「割合Ｍ１－２」と表記する。インスペクターモデルＭ２を用いて算出した第一割合を「割合Ｍ２－１」と表記し、インスペクターモデルＭ２を用いて算出した第二割合を「割合Ｍ２－２」と表記する。インスペクターモデルＭ３を用いて算出した第一割合を「割合Ｍ３－１」と表記し、インスペクターモデルＭ３を用いて算出した第二割合を「割合Ｍ３－２」と表記する。
【０１０７】
たとえば、割合Ｍ１－１は、第１、２、３訓練データセットをインスペクターモデルＭ１に入力した場合に、全訓練データのうち、危険領域４２Ａに含まれる訓練データの割合を示す。割合Ｍ１－２は、運用データセットをインスペクターモデルＭ１に入力した場合に、全運用データのうち、危険領域４２Ａに含まれる運用データの割合を示す。
【０１０８】
割合Ｍ２－１は、第１、２、３訓練データセットをインスペクターモデルＭ２に入力した場合に、全訓練データのうち、危険領域４２Ｂに含まれる訓練データの割合を示す。割合Ｍ２－２は、運用データセットをインスペクターモデルＭ２に入力した場合に、全運用データのうち、危険領域４２Ｂに含まれる運用データの割合を示す。
【０１０９】
割合Ｍ３－１は、第１、２、３訓練データセットをインスペクターモデルＭ３に入力した場合に、全訓練データのうち、危険領域４２Ｃに含まれる訓練データの割合を示す。割合Ｍ３－２は、運用データセットをインスペクターモデルＭ３に入力した場合に、全運用データのうち、危険領域４２Ｃに含まれる運用データの割合を示す。
【０１１０】
情報処理装置は、第一割合と第二割合との差分（差分の絶対値）が閾値以上となった場合に、監視対象の機械学習モデルの精度劣化を検出する。また、情報処理装置は、差分が最も大きい第一割合と第二割合との組を基にして、精度劣化の要因となる分類クラスを特定する。閾値は、予め設定されているものとする。図２１の説明では、閾値を「０．１」とする。
【０１１１】

具体的には、情報処理装置は、割合Ｍ１－１と割合Ｍ１－２との差分の絶対が閾値以上となった場合には、第１クラスが精度劣化の要因と判定する。割合Ｍ２－１と割合Ｍ２－２との差分の絶対が閾値以上となった場合には、第２クラスが精度劣化の要因と判定する。情報処理装置は、割合Ｍ３－１と割合Ｍ３－２との差分の絶対が閾値以上となった場合には、第３クラスが精度劣化の要因と判定する。

【０１１２】

たとえば、割合Ｍ１－１＝０．０９とし、割合Ｍ１－２＝０．３２とすると、割合Ｍ１－１と割合Ｍ１－２との差分の絶対値が「０．２３」となり、閾値以上となる。割合Ｍ２－１＝０．０５とし、割合Ｍ２－２＝０．０５１とすると、割合Ｍ２－１と割合Ｍ２－２との差分の絶対値が「０．０１」となり閾値未満となる。割合Ｍ３－１＝０．００６とし、割合Ｍ３－２＝０．００４とすると、割合Ｍ３－１と割合Ｍ３－２との差分の絶対値が「０．００２」となり、閾値未満となる。この場合には、情報処理装置は、運用データセットのコンセプトドリフトを検知し、精度劣化の要因を、第１クラスとして判定する。

【０１１３】

このように、本実施例２に係る情報処理装置は、３種類以上の分類クラスについて、分類クラス毎に１対他の蒸留を行うことによって、監視対象となる機械学習モデルの精度劣化を検知する。また、情報処理装置は、精度劣化を検知した場合に、インスペクターモデルＭ１～Ｍ３の第一割合と第二割合とを比較することで、どの分類クラスに影響が出ているのかを特定することができる。

【０１１４】

次に、本実施例２に係る情報処理装置の構成について説明する。図２２は、本実施例２に係る情報処理装置の構成を示す機能ブロック図である。図２２に示すように、情報処理装置２００は、通信部２１０と、入力部２２０と、表示部２３０と、記憶部２４０と、制御部２５０とを有する。

【０１１５】

通信部２１０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処理部である。通信部２１０は、通信装置の一例である。後述する制御部２５０は、通信部１１０を介して、外部装置とデータをやり取りする。

【０１１６】

入力部２２０は、情報処理装置２００に対して各種の情報を入力するための入力装置である。入力部２２０は、キーボードやマウス、タッチパネル等に対応する。

【０１１７】

表示部２３０は、制御部２５０から出力される情報を表示する表示装置である。表示部２３０は、液晶ディスプレイ、有機ＥＬディスプレイ、タッチパネル等に対応する。

【０１１８】

記憶部２４０は、教師データ２４１、機械学習モデルデータ２４２、蒸留データテーブル２４３、インスペクターモデルテーブル２４４、運用データテーブル２４５を有する。記憶部１４０は、ＲＡＭ、フラッシュメモリなどの半導体メモリ素子や、ＨＤＤなどの記憶装置に対応する。

【０１１９】

教師データ２４１は、訓練データセット２４１ａと、検証データ２４１ｂを有する。訓練データセット２４１ａは、訓練データに関する各種の情報を保持する。

【０１２０】

図２３は、本実施例２に係る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２３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する。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。本実施例２では、正解ラベルとして、第１クラス、第２クラス、第３クラスのいずれか一つが、訓練データに対応付けられる。

【０１２１】
検証データ２４１ｂは、訓練データセット２４１ａによって学習された機械学習モデルを検証するためのデータである。検証データ２４１ｂに関するその他の説明は、実施例１で説明した検証データ１４１ｂと同様である。
【０１２２】
機械学習モデルデータ２４２は、機械学習モデルのデータである。本実施例２に機械学習モデルは、所定の分類アルゴリズムによって、入力データを、第１クラス、第２クラスまたは第３クラスに分類する機械学習モデルである。分類アルゴリズムは、ＮＮ、ランダムフォレスト、ｋ近傍法、サポートベクターマシン等のうち、いずれの分類アルゴリズムであってもよい。
【０１２３】
本実施例２では、機械学習モデルを、ＮＮとして説明を行う。図２４は、本実施例２に係る機械学習モデルの一例を説明するための図である。図２４に示すように、機械学習モデル５５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、エッジは、重みを持つ。以下の説明では、バイアス値、重みを「パラメータ」と表記する。
【０１２４】
機械学習モデル５５において、入力層５０ａ、隠れ層５０ｂは、図１２で説明した機械学習モデル５０と同様である。機械学習モデル５５は、出力層５０ｃのノード５１ａ，５１ｂ，５１ｃから、各クラスの確率が出力される。たとえば、ノード５１ａから、第１クラスの確率が出力される。ノード５１ｂから、第２クラスの確率が出力される。ノード５１ｃから、第３クラスの確率が出力される。
【０１２５】
蒸留データテーブル２４３は、データセットの各データを、機械学習モデル５５に入力した場合の出力結果を格納するテーブルである。蒸留データテーブルのデータ構造は、実施例１で説明した蒸留データテーブル１４３のデータ構造と同様である。なお、蒸留データテーブル２４３に含まれるソフトターゲットは、第１クラス、第２クラス、第３クラスのうち、いずれかの分類クラスを示すものとする。
【０１２６】
インスペクターモデルテーブル２４４は、ｋＳＶＭによって構築されたインスペクターモデルＭ１，Ｍ２，Ｍ３のデータを格納するテーブルである。各インスペクターモデルＭ１，Ｍ２，Ｍ３に、データを入力すると、符号付きの距離の値が出力される。
【０１２７】
インスペクターモデルＭ１にデータを入力し、符号がプラスであれば、入力したデータは第１クラスに分類される。符号がマイナスであれば、データは、第２クラスまたは第３クラスに分類される。
【０１２８】
インスペクターモデルＭ２にデータを入力し、符号がプラスであれば、入力したデータは第２クラスに分類される。符号がマイナスであれば、データは、第１クラスまたは第３クラスに分類される。
【０１２９】
インスペクターモデルＭ３にデータを入力し、符号がプラスであれば、入力したデータは第３クラスに分類される。符号がマイナスであれば、データは、第１クラスまたは第２クラスに分類される。
【０１３０】
運用データテーブル２４５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。運用データテーブル２４５のデータ構造は、実施例１で説明した運用データテーブル１４５のデータ構造と同様である。

【０１３１】
図２２の説明に戻る。制御部２５０は、学習部２５１と、作成部２５２と、検出部２５３と、予測部２５４とを有する。制御部２５０は、ＣＰＵやＭＰＵなどによって実現できる。また、制御部２５０は、ＡＳＩＣやＦＰＧＡなどのハードワイヤードロジックによっても実現できる。
【０１３２】
学習部２５１は、訓練データセット２４１ａを取得し、訓練データセット２４１ａを基にして、機械学習モデル５５のパラメータを学習する処理部である。たとえば、学習部２５１は、訓練データセット２４１ａの訓練データを、機械学習モデル５５の入力層に入力した場合、出力層の各ノードの出力結果が、入力した訓練データの正解ラベルに近づくように、機械学習モデル５５のパラメータを更新する（誤差逆伝播法による学習）。学習部２５１は、訓練データセット２４１ａに含まれる各訓練データについて、上記処理を繰り返し実行する。また、学習部２５１は、検証データ２４１ｂを用いて、機械学習モデル５５の検証を行ってもよい。学習部２５１は、学習済みの機械学習モデル５５のデータ（機械学習モデルデータ２４２）を、記憶部２４０に登録する。機械学習モデル５５は、「運用モデル」の一例である。
【０１３３】
図２５は、本実施例２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２４１ａ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、×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」とし、丸印の訓練データに対応する正解ラベルを「第３クラス」とする。
【０１３４】
たとえば、訓練データセット２４１ａによって、機械学習モデル５５を学習すると、特徴空間３０は、決定境界３６によって、モデル適用領域３６Ａと、モデル適用領域３６Ｂと、モデル適用領域３６Ｃとに分類される。たとえば、機械学習モデル５５が、ＮＮである場合、機械学習モデル５５に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、第３クラスの確率がそれぞれ出力される。第１クラスの確率が、他の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、他の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。第３クラスの確率が、他のクラスよりも大きい場合には、データは、第３クラスに分類される。
【０１３５】
作成部２５２は、機械学習モデル５５の知識蒸留を基にして、インスペクターモデルＭ１，Ｍ２，Ｍ３を作成する処理部である。たとえば、作成部２５２は、「モデル適用領域３６Ａ」と「モデル適用領域３６Ｂ，３６Ｃ」との決定境界（図２１の決定境界４１Ａに相当）を学習した、インスペクターモデルＭ１を作成する。このインスペクターモデルＭ１にデータ（訓練データまたは運用データ）を入力すると、決定境界４１Ａとデータとの距離（符号付きの距離の値）が出力される。
【０１３６】
作成部２５２は、「モデル適用領域３６Ｂ」と「モデル適用領域３６Ａ，３６Ｃ」との決定境界（図２１の決定境界４１Ｂに相当）を学習した、インスペクターモデルＭ２を作成する。このインスペクターモデルＭ２にデータ（訓練データまたは運用データ）を入力すると、決定境界４１Ｂとデータとの距離（符号付きの距離の値）が出力される。
【０１３７】
作成部２５２は、「モデル適用領域３６Ｃ」と「モデル適用領域３６Ａ，３６Ｂ」との決定境界（図２１の決定境界４１Ｃに相当）を学習した、インスペクターモデルＭ３を作成する。このインスペクターモデルＭ３にデータ（訓練データまたは運用データ）を入力すると、決定境界４１Ｃとデータとの距離（符号付きの距離の値）が出力される。

【０１３８】
図２６は、インスペクターモデルの決定境界および危険領域の一例を示す図である。図２６では、一例として、インスペクターモデルＭ２の決定境界および危険領域４２Ｂを示す。インスペクターモデルＭ１，Ｍ３に係る決定境界および危険領域の図示を省略する。
【０１３９】
作成部２５２は、蒸留データテーブル２４３を生成する処理、インスペクターモデルテーブル２４４を作成する処理を実行する。
【０１４０】
まず、作成部２５２が、蒸留データテーブル２４３を生成する処理について説明する。作成部２５２は、機械学習モデルデータ２４２を用いて、機械学習モデル５５を実行し、特徴空間上の各データを、機械学習モデル５５に入力する。これにより、特徴空間の各データが、第１クラス、第２クラス、第３クラスのうち、いずれの分類クラスに分類されるのかを特定する。かかる処理を実行することで、作成部２５２は、特徴空間をモデル適用領域３６Ａと、モデル適用領域３６Ｂ，モデル適用領域３６Ｃとに分類し、決定境界３６を特定する。
【０１４１】
作成部２５２は、特徴空間３０上において「グリッド」を配置する。グリッドの幅は、予め設定されているものとする。作成部２５２は、グリッドの交点座標のデータを選択し、選択したデータを、機械学習モデル５５に出力することで、選択したデータに対応するソフトターゲットを算出する。作成部２５２は、選択したデータ（入力データ）と、ソフトターゲットとを対応付けて、蒸留データテーブル２４３に登録する。作成部２５２は、グリッドの各交点座標のデータについても、上記処理を繰り返し実行することで、蒸留データテーブル２４３を生成する。
【０１４２】
続いて、作成部２５２が、インスペクターモデルテーブル２４４を作成する処理について説明する。作成部２５２は、蒸留データテーブル２４３に登録された入力データと、ソフトターゲットとの関係を基にして、ｋＳＶＭによって構築されたインスペクターモデルＭ１～Ｍ３を作成する。作成部２５２は、作成したインスペクターモデルＭ１～Ｍ３のデータを、インスペクターモデルテーブル２４４に登録する。
【０１４３】
作成部２５２が、「インスペクターモデルＭ１」を作成する処理の一例について説明する。作成部２５２は、蒸留データテーブル２４３に格納された各入力データを、再生核ヒルベルト空間に射影する。作成部２５２は、再生核ヒルベルト空間に含まれる第１クラスの入力データのうち、決定境界４１Ａに最も近い入力データを、第１サポートベクトルとして選択する。作成部１５２は、再生核ヒルベルト空間に含まれる第２クラスまたは第３クラスの入力データのうち、決定境界４１Ａに最も近い入力データを、第２サポートベクトルとして選択する。作成部２５２は、第１サポートベクトルと、第２サポートベクトルとの中間を通る決定境界４１Ａを特定することで、インスペクターモデルＭ１のハイパーパラメータを特定する。再生核ヒルベルト空間において、決定境界４１Ａは直線となり、決定境界４１Ａからの距離がm_{M1}となる領域を、危険領域４２Ａに設定する。距離m_{M1}は、決定境界４１Ａと、第１サポートベクトル（第２サポートベクトル）との距離である。
【０１４４】
作成部２５２が、「インスペクターモデルＭ２」を作成する処理の一例について説明する。作成部２５２は、蒸留データテーブル２４３に格納された各入力データを、再生核ヒルベルト空間に射影する。作成部２５２は、再生核ヒルベルト空間に含まれる第２クラスの入力データのうち、決定境界４１Ｂに最も近い入力データを、第３サポートベクトルとして選択する。作成部２５２は、再生核ヒルベルト空間に含まれる第１クラスまたは第３クラスの入力データのうち、決定境界４１Ｂに最も近い入力データを、第４サポートベクトルとして選択する。作成部２５２は、第３サポートベクトルと、第４サポートベクトルとの中間を通る決定境界４１Ｂを特定することで、インスペクターモデルＭ２のハイパー

パラメータを特定する。再生核ヒルベルト空間において、決定境界４１Ｂは直線となり、決定境界４１Ｂからの距離がm_{M2}となる領域を、危険領域４２Ｂに設定する。距離m_{M2}は、決定境界４１Ｂと、第３サポートベクトル（第４サポートベクトル）との距離である。

【０１４５】

作成部２５２が、「インスペクターモデルＭ３」を作成する処理の一例について説明する。作成部２５２は、蒸留データテーブル２４３に格納された各入力データを、再生核ヒルベルト空間に射影する。作成部２５２は、再生核ヒルベルト空間に含まれる第３クラスの入力データのうち、決定境界４１Ｃに最も近い入力データを、第５サポートベクトルとして選択する。作成部２５２は、再生核ヒルベルト空間に含まれる第１クラスまたは第２クラスの入力データのうち、決定境界４１Ｃに最も近い入力データを、第６サポートベクトルとして選択する。作成部２５２は、第５サポートベクトルと、第６サポートベクトルとの中間を通る決定境界４１Ｃを特定することで、インスペクターモデルＭ３のハイパーパラメータを特定する。再生核ヒルベルト空間において、決定境界４１Ｃは直線となり、決定境界４１Ｃからの距離がm_{M3}となる領域を、危険領域４２Ｃに設定する。距離m_{M3}は、決定境界４１Ｃと、第５サポートベクトル（第６サポートベクトル）との距離である。

【０１４６】

検出部２５３は、インスペクターモデルＭ１～Ｍ３を実行して、機械学習モデル５５の精度劣化を検出する処理部である。また、検出部２５３は、機械学習モデル５５の精度劣化を検出した場合、精度劣化の要因となる分類クラスを特定する。

【０１４７】

検出部２５３は、インスペクターモデルＭ１～Ｍ３に訓練データセット２４１ａをそれぞれ入力することで、各第一割合（割合Ｍ１－１、割合Ｍ２－１、割合Ｍ３－１）を算出する。

【０１４８】

検出部２５３は、訓練データを、インスペクターモデルＭ１に入力すると、特徴空間上の決定境界４１Ａと訓練データとの距離が出力される。検出部２５３は、決定境界４１Ａと訓練データとの距離が距離m_{M1}未満である場合、かかる訓練データが危険領域４２Ａに含まれると判定する。検出部２５３は、各訓練データに対して、上記処理を繰り返し実行し、全訓練データのうち、危険領域４２Ａに含まれる訓練データの数を特定し、割合Ｍ１－１を算出する。

【０１４９】

検出部２５３は、訓練データを、インスペクターモデルＭ２に入力すると、特徴空間上の決定境界４１Ｂと訓練データとの距離が出力される。検出部２５３は、決定境界４１Ｂと訓練データとの距離が距離m_{M2}未満である場合、かかる訓練データが危険領域４２Ｂに含まれると判定する。検出部２５３は、各訓練データに対して、上記処理を繰り返し実行し、全訓練データのうち、危険領域４２Ｂに含まれる訓練データの数を特定し、割合Ｍ２－１を算出する。

【０１５０】

検出部２５３は、訓練データを、インスペクターモデルＭ３に入力すると、特徴空間上の決定境界４１Ｃと訓練データとの距離が出力される。検出部２５３は、決定境界４１Ｃと訓練データとの距離が距離m_{M3}未満である場合、かかる訓練データが危険領域４２Ｃに含まれると判定する。検出部２５３は、各訓練データに対して、上記処理を繰り返し実行し、全訓練データのうち、危険領域４２Ｃに含まれる訓練データの数を特定し、割合Ｍ３－１を算出する。

【０１５１】

検出部２５３は、インスペクターモデルＭ１～Ｍ３に運用データセットをそれぞれ入力することで、各第二割合（割合Ｍ１－２、割合Ｍ２－２、割合Ｍ３－２）を算出する。

【０１５２】

検出部２５３は、運用データを、インスペクターモデルＭ１に入力すると、特徴空間上の決定境界４１Ａと運用データとの距離が出力される。検出部２５３は、決定境界４１Ａ

と訓練データとの距離が距離m_{M1}未満である場合、かかる運用データが危険領域４２Ａに含まれると判定する。検出部２５３は、各運用データに対して、上記処理を繰り返し実行し、全運用データのうち、危険領域４２Ａに含まれる運用データの数を特定し、割合Ｍ１－２を算出する。

【０１５３】

検出部２５３は、運用データを、インスペクターモデルＭ２に入力すると、特徴空間上の決定境界４１Ｂと運用データとの距離が出力される。検出部２５３は、決定境界４１Ｂと運用データとの距離が距離m_{M2}未満である場合、かかる運用データが危険領域４２Ｂに含まれると判定する。検出部２５３は、各運用データに対して、上記処理を繰り返し実行し、全運用データのうち、危険領域４２Ｂに含まれる運用データの数を特定し、割合Ｍ２－１を算出する。

【０１５４】

検出部２５３は、運用データを、インスペクターモデルＭ３に入力すると、特徴空間上の決定境界４１Ｃと運用データとの距離が出力される。検出部２５３は、決定境界４１Ｃと運用データとの距離が距離m_{M3}未満である場合、かかる運用データが危険領域４２Ｃに含まれると判定する。検出部２５３は、各運用データに対して、上記処理を繰り返し実行し、全運用データのうち、危険領域４２Ｃに含まれる運用データの数を特定し、割合Ｍ３－１を算出する。

【０１５５】

検出部２５３は、対応する第一割合と第二割合とを比較して、第一割合に対して第二割合が変化した場合に、コンセプトドリフトが発生したと判定し、機械学習モデル５５の精度劣化を検出する。たとえば、検出部２５３は、第一割合と第二割合との差分の絶対値が閾値以上である場合に、コンセプトドリフトが発生したと判定する。

【０１５６】

ここで、対応する第一割合と第二割合との組を、割合Ｍ１－１と割合Ｍ１－２との組、割合Ｍ２－１と割合Ｍ２－２との組、割合Ｍ３－１と割合Ｍ３－２との組とする。

【０１５７】

また、検出部２５３は、割合Ｍ１－１と割合Ｍ１－２との差分の絶対値が閾値以上となる場合に、精度劣化の要因となるクラスを「第１クラス」と判定する。検出部２５３は、割合Ｍ２－１と割合Ｍ２－２との差分の絶対値が閾値以上となる場合に、精度劣化の要因となるクラスを「第２クラス」と判定する。検出部２５３は、割合Ｍ３－１と割合Ｍ３－２との差分の絶対値が閾値以上となる場合に、精度劣化の要因となるクラスを「第３クラス」と判定する。

【０１５８】

検出部２５３は、上記処理によって、機械学習モデル５５の精度劣化を検出した場合、精度劣化を検知した旨と、精度劣化の要因となる分類クラスの情報を、表示部２３０に出力して表示する。また、検出部２５３は、精度劣化を検知した旨と、精度劣化の要因となる分類クラスの情報を、外部装置に送信してもよい。

【０１５９】

検出部２５３は、機械学習モデル５５の精度劣化を検出しない場合には、精度劣化を検出していない旨の情報を予測部２５４に出力する。

【０１６０】

予測部２５４は、機械学習モデル５５の精度劣化が検出されていない場合、機械学習モデル５５を実行して、運用データセットを入力し、各運用データの分類クラスを予測する処理部である。予測部２５４は、予測結果を、表示部２３０に出力して表示させてもよいし、外部装置に送信してもよい。

【０１６１】

次に、本実施例２に係る情報処理装置２００の処理手順の一例について説明する。図２７は、本実施例２に係る情報処理装置の処理手順を示すフローチャートである。図２７に示すように、情報処理装置２００の学習部２５１は、訓練データセット２４１ａを基にし

て、機械学習モデル５５を学習する（ステップＳ２０１）。
【０１６２】
情報処理装置２００の作成部２５２は、知識蒸留を用いて、蒸留データテーブル２４３を生成する（ステップＳ２０２）。情報処理装置２００の作成部２５２は、蒸留データテーブル２４３を基にして、複数のインスペクターモデルＭ１～Ｍ３を作成する（ステップＳ２０３）。
【０１６３】
情報処理装置２００の検出部２５３は、訓練データセットの各訓練データをインスペクターモデルＭ１～Ｍ３にそれぞれ入力し、各第一割合（割合Ｍ１－１、割合Ｍ２－１、割合Ｍ３－１）を算出する（ステップＳ２０４）。
【０１６４】
検出部２５３は、運用データセットの各運用データをインスペクターモデルＭ１～Ｍ３にそれぞれ入力し、各第二割合（割合Ｍ１－２、割合Ｍ２－２、割合Ｍ３－２）を算出する（ステップＳ２０５）。
【０１６５】
検出部２５３は、各第一割合と各第二割合とを基にして、コンセプトドリフトが発生したか否かを判定する（ステップＳ２０６）。情報処理装置２００は、コンセプトドリフトが発生した場合には（ステップＳ２０７，Ｙｅｓ）、ステップＳ２０８に移行する。一方、情報処理装置２００は、コンセプトドリフトが発生していない場合には（ステップＳ２０７，Ｎｏ）、ステップＳ２０９に移行する。
【０１６６】
ステップＳ２０８以降の処理について説明する。学習部２５１は、新たな訓練データセットによって、機械学習モデル５５を再学習し（ステップＳ２０８）、ステップＳ２０２に移行する。
【０１６７】
ステップＳ２０９以降の処理について説明する。情報処理装置２００の予測部２５４は、運用データセットを、機械学習モデル５５に入力し、各運用データの分類クラスを予測する（ステップＳ２０９）。予測部２５４は、予測結果を出力する（ステップＳ２１０）。
【０１６８】
次に、本実施例２に係る情報処理装置２００の効果について説明する。情報処理装置２００は、３種類以上の分類クラスについて、分類クラス毎に１対他の蒸留を行うことによって、監視対象となる機械学習モデルの精度劣化を検知する。また、情報処理装置２００は、精度劣化を検知した場合に、どの分類クラスに影響が出ているのかを特定することができる。
【０１６９】
たとえば、分類クラスが３つ以上の場合には、決定境界からの距離のみでは、どの方向に運用データがコンセプトドリフトしているかを特定することができない。これに対して、１対他のクラスの分類モデル（複数のインスペクターモデルＭ１～Ｍ３）を作成することで、どの方向にコンセプトドリフトしているのかを特定でき、どの分類クラスに影響が出ているのかを特定することができる。
【実施例３】
【０１７０】
本実施例３に係る情報処理装置は、運用データセットに含まれる一つの運用データ毎に、コンセプトドリフト（精度劣化の要因）が発生しているか否かを判定する。以下の説明では、データセットに含まれる一つのデータ（訓練データまたは運用データ）を、「インスタンス」と表記する。
【０１７１】
図２８は、本実施例３に係る情報処理装置の処理を説明するための図である。本実施例３に係る情報処理装置は、実施例１の情報処理装置１００と同様にして、知識蒸留を用いて、インスペクターモデルを作成する。インスペクターモデルによって学習した決定境界

を、決定境界６０とする。情報処理装置は、特徴空間上のインスタンスと、決定境界６０との距離を基にして、精度劣化の要因となるインスタンスとして検出する。

【０１７２】

たとえば、図２８において、運用データセット６１に含まれるインスタンス毎に、確信度は異なる。たとえば、インスタンス６１ａと、決定境界６０との距離はｄａである。インスタンス６１ｂと、決定境界６０との距離はｄｂである。距離ｄａは、距離ｄｂよりも小さいため、インスタンス６１ａは、インスタンス６１ｂよりも、精度劣化の要因となり得る。

【０１７３】

ここで、決定境界とインスタンスとの距離はスカラー値であり、運用データセット毎に大きさが変化するため、どれくらいの決定境界からの距離が危ないのかを特定するための閾値を設定することが難しい。このため、情報処理装置は、決定境界からの距離を確率値へと変換し、変換した確率値を確信度として取り扱う。これによって、確信度は、運用データセットによらず、「０～１」の値をとる。

【０１７４】

たとえば、情報処理装置は、式（２）に基づいて、確信度を算出する。式（２）に示す例では、あるインスタンスが第１クラスである確率を示すものである。インスタンスの特徴量を「ｘ」とし、決定境界とインスタンスとの距離を「ｆ（ｘ）」とする。「Ａ」および「Ｂ」は、訓練データセットから学習されるハイパーパラメータである。

【０１７５】

Ｐ（ｙ＝１｜ｘ）＝１／（１＋ｅｘｐ（Ａｆ（ｘ）＋Ｂ））・・・（２）

【０１７６】

情報処理装置は、式（２）に基づいて、運用データセットのインスタンスの確信度を算出し、確信度が予め設定された閾値未満である場合に、かかるインスタンスを、精度劣化の要因として特定する。これによって、運用データセットによらず、確信度を「０～１」の範囲で算出でき、精度劣化の要因となるインスタンスを適切に特定する。

【０１７７】

ところで、本実施例３に係る情報処理装置は、更に、次の処理を実行して、監視対象となる機械学習モデルの精度劣化を検出してもよい。情報処理装置は、訓練データセットの各訓練データを、インスペクターモデルに入力して、各訓練データと決定境界６０との距離をそれぞれ算出し、各距離の平均値を「第１の距離」として特定する。

【０１７８】

情報処理装置は、運用データセットの各運用データを、インスペクターモデルに入力して、各運用データと決定境界６０との距離をそれぞれ算出し、各距離の平均値を「第２の距離」として特定する。

【０１７９】

情報処理装置は、第１の距離と、第２の距離との差分が予め設定された閾値以上の場合に、コンセプトドリフトが発生したものとして、機械学習モデルの精度劣化を検出する。

【０１８０】

上記のように、本実施例３に係る情報処理装置は、決定境界６０と、インスタンスとの距離を算出することで、精度劣化の要因となるインスタンスを特定することが可能になる。また、訓練データセットの各インスタンスに基づく第１の距離と、運用データセットの各インスタンスに基づく第２の距離とを利用することで、機械学習モデルの精度劣化を検出することもできる。

【０１８１】

次に、本実施例３に係る情報処理装置の構成の一例について説明する。図２９は、本実施例３に係る情報処理装置の構成を示す機能ブロック図である。図２９に示すように、この情報処理装置３００は、通信部３１０と、入力部３２０と、表示部３３０と、記憶部３４０と、制御部３５０とを有する。

【０１８２】

通信部３１０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処理部である。通信部３１０は、通信装置の一例である。後述する制御部３５０は、通信部３１０を介して、外部装置とデータをやり取りする。

【０１８３】

入力部３２０は、情報処理装置３００に対して各種の情報を入力するための入力装置である。入力部３２０は、キーボードやマウス、タッチパネル等に対応する。

【０１８４】

表示部３３０は、制御部３５０から出力される情報を表示する表示装置である。表示部３３０は、液晶ディスプレイ、有機ＥＬディスプレイ、タッチパネル等に対応する。

【０１８５】

記憶部３４０は、教師データ３４１、機械学習モデルデータ３４２、蒸留データテーブル３４３、インスペクターモデルデータ３４４、運用データテーブル３４５を有する。記憶部３４０は、ＲＡＭ、フラッシュメモリなどの半導体メモリ素子や、ＨＤＤなどの記憶装置に対応する。

【０１８６】

教師データ３４１は、訓練データセット３４１ａと、検証データ３４１ｂを有する。訓練データセット３４１ａは、訓練データに関する各種の情報を保持する。訓練データセット３４１ａのデータ構造に関する説明は、実施例１で説明した訓練データセット１４１ａのデータ構造に関する説明と同様である。

【０１８７】

検証データ３４１ｂは、訓練データセット３４１ａによって学習された機械学習モデルを検証するためのデータである。

【０１８８】

機械学習モデルデータ３４２は、機械学習モデルのデータである。機械学習モデルデータ３４２に関する説明は、実施例１で説明した機械学習モデルデータ１４２に関する説明と同様である。本実施例３では、監視対象の機械学習モデルを、機械学習モデル５０として説明を行う。なお、機械学習モデルの分類アルゴリズムは、ＮＮ、ランダムフォレスト、ｋ近傍法、サポートベクターマシン等のうち、いずれの分類アルゴリズムであってもよい。

【０１８９】

蒸留データテーブル３４３は、データセットの各データを、機械学習モデル５０に入力した場合の出力結果（ソフトターゲット）を格納するテーブルである。蒸留データテーブル３４３のデータ構造に関する説明は、実施例１で説明した蒸留データテーブル１４３のデータ構造に関する説明と同様である。

【０１９０】

インスペクターモデルデータ３４４は、ｋＳＶＭによって構築されたインスペクターモデルのデータである。インスペクターモデルデータ３４４に関する説明は、実施例１で説明したインスペクターモデルデータ１４４に関する説明と同様である。

【０１９１】

運用データテーブル３４５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。運用データテーブル３４５のデータ構造に関する説明は、実施例１で説明した運用データテーブル１４５に関する説明と同様である。

【０１９２】

制御部３５０は、学習部３５１と、作成部３５２と、検出部３５３と、予測部３５４とを有する。制御部３５０は、ＣＰＵやＭＰＵなどによって実現できる。また、制御部３５０は、ＡＳＩＣやＦＰＧＡなどのハードワイヤードロジックによっても実現できる。

【０１９３】

学習部３５１は、訓練データセット３４１ａを取得し、訓練データセット３４１ａを基にして、機械学習モデル５０のパラメータを学習する処理部である。学習部３５１の処理に関する説明は、実施例１で説明した学習部１５１の処理に関する説明と同様である。

【０１９４】
作成部３５２は、機械学習モデル５０の知識蒸留を基にして、モデル適用領域３１Ａとモデル適用領域３１Ｂとの決定境界３１を学習した、インスペクターモデルを作成する処理部である。作成部３５２が、インスペクターモデルを作成する処理は、実施例１で説明した作成部１５２が、インスペクターモデルを作成する処理と同様である。
【０１９５】
なお、作成部３５２は、訓練データセット３４１ａの各訓練データおよび正解ラベルを基にして、式（２）で説明したハイパーパラメータＡ，Ｂを学習する。たとえば、作成部３５２は、正解ラベル「第１クラス」に対応する訓練データの特徴量ｘを、式（２）に入力した場合の値が１に近づくように、ハイパーパラメータＡ、Ｂを調整する。作成部３５２は、正解ラベル「第２クラス」に対応する訓練データの特徴量ｘを、式（２）に入力した場合の値が０に近づくように、ハイパーパラメータＡ、Ｂを調整する。作成部３５２は、各訓練データを用いて、上記処理を繰り返し実行することで、ハイパーパラメータＡ，Ｂを学習する。作成部３５２は、学習したハイパーパラメータＡ，Ｂのデータを、検出部３５３に出力する。
【０１９６】
検出部３５３は、機械学習モデル５０の精度劣化の要因となるインスタンスを検出する処理部である。検出部３５３は、インスペクターモデル３５を実行する。検出部３５３は、運用データセットに含まれるインスタンス（運用データ）を選択し、選択したインスタンスを、インスペクターモデル３５に入力することで、決定境界３１と、インスタンスとの距離を特定する。また、検出部３５３は、特定した距離ｆ（ｘ）を、式（２）に入力することで、選択したインスタンスの確信度を算出する。
【０１９７】
検出部３５３は、確信度が閾値未満である場合に、選択したインスタンスを、精度劣化の要因となるインスタンスとして検出する。検出部３５３は、運用データセットに含まれる各運用データについて、上記処理を繰り返し実行することで、精度劣化の要因となる運用データを検出する。
【０１９８】
検出部３５３は、精度劣化の要因となる各インスタンス（運用データ）のデータを、表示部３３０に出力して表示させてもよいし、外部装置に送信してもよい。
【０１９９】
ところで、検出部３５３は、更に、次の処理を実行して、監視対象となる機械学習モデル５０の精度劣化を検出してもよい。検出部３５３は、訓練データセット３４１ａの各訓練データを、インスペクターモデル３５に入力して、各訓練データと決定境界６０との距離をそれぞれ算出し、各距離の平均値を「第１の距離」として特定する。
【０２００】
検出部３５３は、運用データテーブル３４５から運用データセットを選択する。検出部３５３は、運用データセットの各運用データを、インスペクターモデル３５に入力して、各運用データと決定境界６０との距離をそれぞれ算出し、各距離の平均値を「第２の距離」として特定する。
【０２０１】
検出部３５３は、第１の距離と、第２の距離との差分が予め設定された閾値以上の場合に、コンセプトドリフトが発生したものとして、機械学習モデル５０の精度劣化を検出する。検出部３５３は、時間経過に伴って追加され各運用データセットについて、上記処理を繰り返し実行し、機械学習モデル５０の精度劣化を検出する。
【０２０２】
検出部３５３は、機械学習モデル５０の精度劣化を検出した場合には、精度劣化を検出した旨の情報を、表示部３３０に表示してもよいし、外部装置（図示略）に、精度劣化を検出した旨を通知してもよい。検出部３５３は、精度劣化を検出した根拠となる運用データセットのデータ識別情報を、表示部３３０に出力して表示させてもよい。また、検出部

３５３は、精度劣化を検出した旨を学習部３５１に通知して、機械学習モデルデータ３４２を再学習させてもよい。

【０２０３】

予測部３５４は、機械学習モデル５０の精度劣化が検出されていない場合、機械学習モデル５０を実行して、運用データセットを入力し、各運用データの分類クラスを予測する処理部である。予測部３５４は、予測結果を、表示部３３０に出力して表示させてもよいし、外部装置に送信してもよい。

【０２０４】

次に、本実施例３に係る情報処理装置３００の処理手順の一例について説明する。図３０は、本実施例３に係る情報処理装置の処理手順を示すフローチャートである。図３０に示すように、情報処理装置３００の学習部３５１は、訓練データセット３４１ａを基にして、機械学習モデル５０を学習する（ステップＳ３０１）。

【０２０５】

情報処理装置３００の作成部３５２は、知識蒸留を用いて、蒸留データテーブル３４３を生成する（ステップＳ３０２）。作成部３５２は、蒸留データテーブル３４３を基にして、インスペクターモデルを作成する（ステップＳ３０３）。作成部３５２は、訓練データセット３４１ａを用いて、式（２）のハイパーパラメータＡ，Ｂを学習する（ステップＳ３０４）。

【０２０６】

情報処理装置３００の検出部３５３は、運用データセットのインスタンスを選択する（ステップＳ３０５）。検出部３５３は、選択したインスタンスをインスペクターモデルに入力し、決定境界とインスタンスとの距離を算出する（ステップＳ３０６）。検出部３５３は、インスタンスの確信度を算出する（ステップＳ３０７）。

【０２０７】

検出部３５３は、インスタンスの確信度が閾値未満でない場合には（ステップＳ３０８，Ｎｏ）、ステップＳ３１０に移行する。一方、検出部３５３は、インスタンスの確信度が閾値未満である場合には（ステップＳ３０８，Ｙｅｓ）、ステップＳ３０９に移行する。

【０２０８】

検出部３５３は、選択したインスタンスを、精度劣化の要因として特定する（ステップＳ３０９）。情報処理装置３００は、全てのインスタンスを選択していない場合には（ステップＳ３１０，Ｎｏ）、ステップＳ３１２に移行する。情報処理装置３００は、全てのインスタンスを選択した場合には（ステップＳ３１０，Ｙｅｓ）、ステップＳ３１１に移行する。検出部３５３は、精度劣化の要因として特定したインスタンスを出力する（ステップＳ３１１）。

【０２０９】

ステップＳ３１２以降の処理について説明する。検出部３５３は、運用データセットから次のインスタンスを選択し（ステップＳ３１２）、ステップＳ３０６に移行する。

【０２１０】

次に、本実施例３に係る情報処理装置３００の効果について説明する。情報処理装置３００は、知識蒸留を用いてインスペクターモデルを学習し、特徴空間上のインスタンスと、決定境界６０との距離を確信度に変換する。確信度に変換することにより、情報処理装置３００は、運用データセットによらず、精度劣化の要因となるインスタンスを検出することができる。

【０２１１】

情報処理装置３００は、訓練データセットの各インスタンスに基づく第１の距離と、運用データセットの各インスタンスに基づく第２の距離とを利用することで、機械学習モデルの精度劣化を検出することもできる。

【０２１２】

次に、本実施例に示した情報処理装置１００（２００，３００）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３１は、本実施例に係る

情報処理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。

【０２１３】

図３１に示すように、コンピュータ４００は、各種演算処理を実行するＣＰＵ４０１と、ユーザからのデータの入力を受け付ける入力装置４０２と、ディスプレイ４０３とを有する。また、コンピュータ４００は、記憶媒体からプログラム等を読み取る読み取り装置４０４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４０５とを有する。コンピュータ４００は、各種情報を一時記憶するＲＡＭ４０６と、ハードディスク装置４０７とを有する。そして、各装置４０１～４０７は、バス４０８に接続される。

【０２１４】

ハードディスク装置４０７は、学習プログラム４０７ａ、作成プログラム４０７ｂ、検出プログラム４０７ｃ、予測プログラム４０７ｄを有する。ＣＰＵ４０１は、学習プログラム４０７ａ、作成プログラム４０７ｂ、検出プログラム４０７ｃ、予測プログラム４０７ｄを読み出してＲＡＭ４０６に展開する。

【０２１５】

学習プログラム４０７ａは、学習プロセス４０６ａとして機能する。作成プログラム４０７ｂは、作成プロセス４０６ｂとして機能する。検出プログラム４０７ｃは、検出プロセス４０６ｃとして機能する。予測プログラム４０７ｄは、予測プロセス４０６ｄとして機能する。

【０２１６】

学習プロセス４０６ａの処理は、学習部１５１，２５１，３５１の処理に対応する。作成プロセス４０６ｂの処理は、作成部１５２，２５２，３５２の処理に対応する。検出プロセス４０６ｃの処理は、検出部１５３，２５３，３５３の処理に対応する。予測プロセス４０６ｄは、予測部１５４，２５４，３５４の処理に対応する。

【０２１７】

なお、各プログラム４０７ａ～４０７ｄついては、必ずしも最初からハードディスク装置４０７に記憶させておかなくてもよい。例えば、コンピュータ４００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４００が各プログラム４０７ａ～４０７ｄを読み出して実行するようにしてもよい。

【符号の説明】

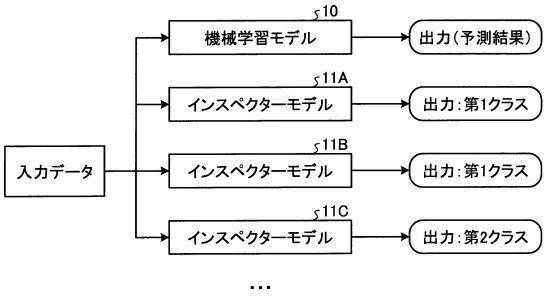
【０２１８】

１００，２００，３００　情報処理装置
１１０，２１０，３１０　通信部
１２０，２２０，３２０　入力部
１３０，２３０，３３０　表示部
１４０，２４０，３４０　記憶部
１４１，２４１，３４１　教師データ
１４１ａ，２４１ａ，３４１ａ　訓練データセット
１４１ｂ，２４１ｂ，３４１ｂ　検証データ
１４２，２４２，３４２　機械学習モデルデータ
１４３，２４３，３４３　蒸留データテーブル
１４４，３４４　インスペクターモデルデータ
１４５，２４５，３４５　運用データテーブル
１５０，２５０，３５０　制御部
１５１，２５１，３５１　学習部
１５２，２５２，３５２　作成部
１５３，２５３，３５３　検出部
１５４，２５４，３５４　予測部

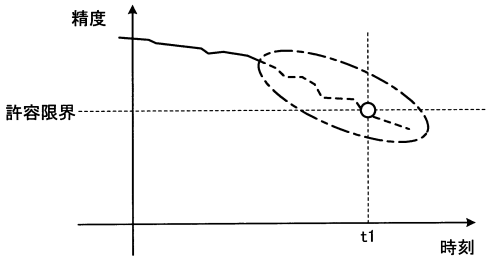
２４４　インスペクターモデルテーブル

【図面】

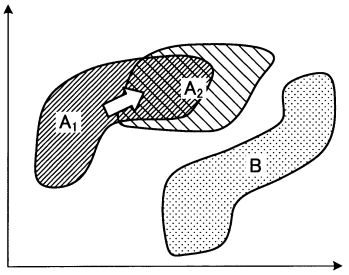
【図１】



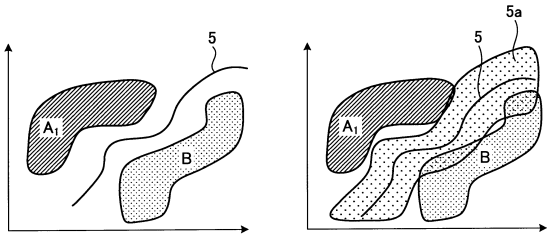
【図２】



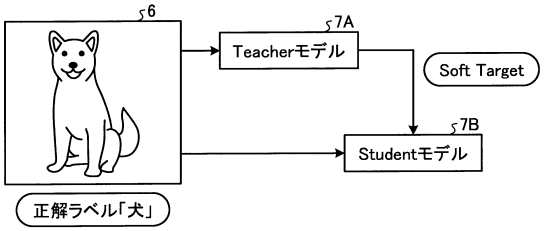
【図３】



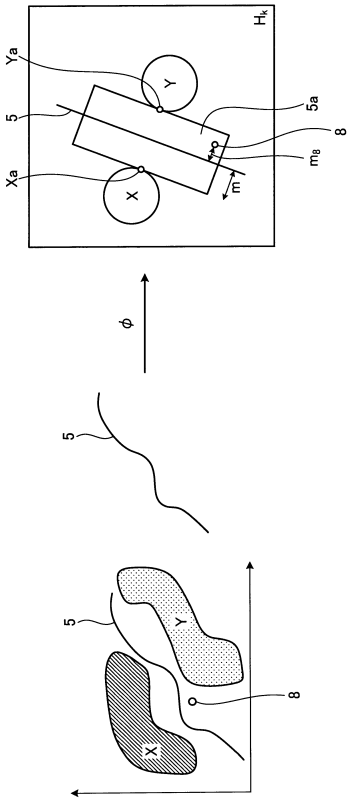
【図４】



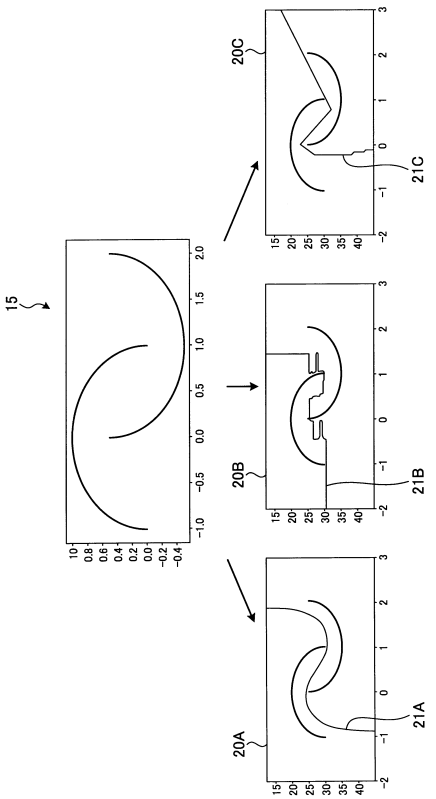
【図５】


6
7A
Teacherモデル
Soft Target
7B
Studentモデル
正解ラベル「犬」

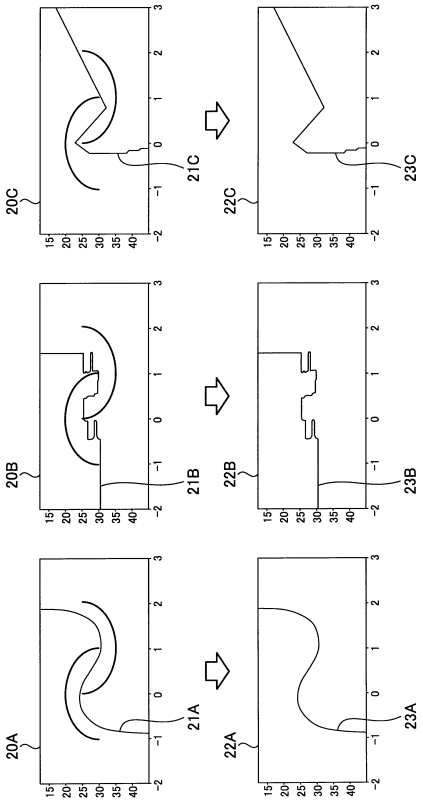
【図６】


Ya
5
Xa
Y
X
H_k
5a
8
m_8
E
φ
5
5
Y
8
X

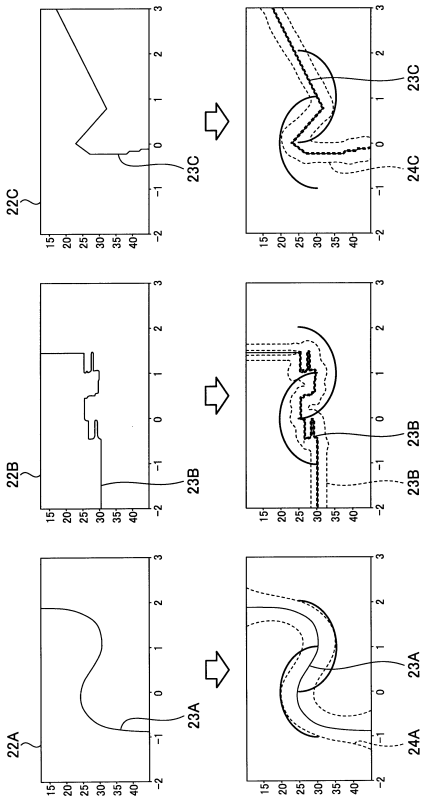
【図７】


15
20C
21C
20B
21B
20A
21A

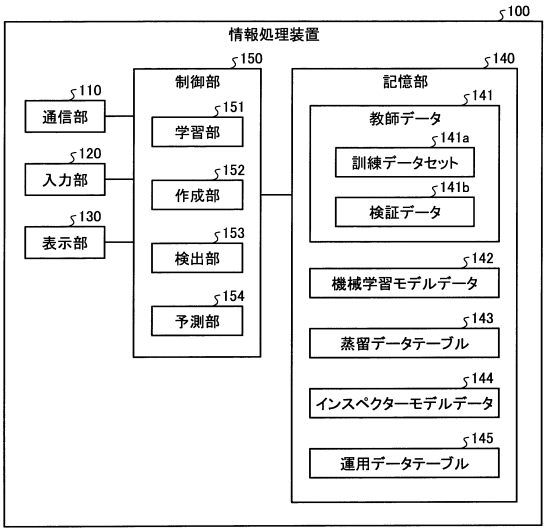
【図８】


20C
21C
22C
23C
20B
21B
22B
23B
20A
21A
22A
23A

【図９】



【図１０】

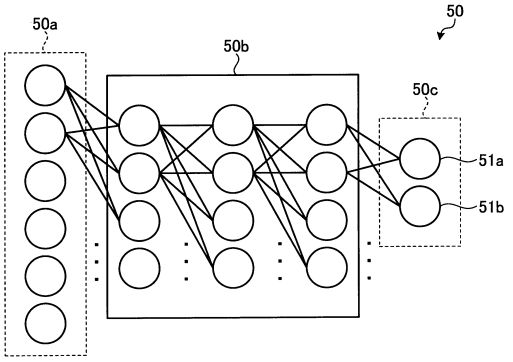


【図１１】

141a

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１２】



【図１３】

143

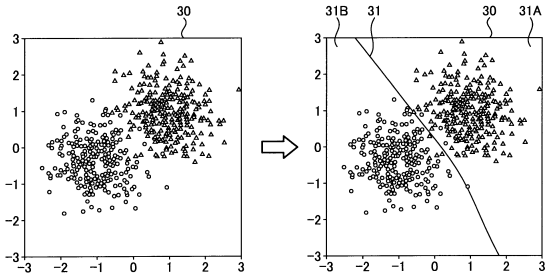
レコード番号	入力データ	ソフトターゲット
1001	レコード番号「1001」 入力データ	レコード番号「1001」入力データを 機械学習モデルに入力した際の出力結果
1002	レコード番号「1002」 入力データ	レコード番号「1002」入力データを 機械学習モデルに入力した際の出力結果
1003	レコード番号「1003」 入力データ	レコード番号「1003」入力データを 機械学習モデルに入力した際の出力結果
…	…	…
1050	レコード番号「1050」 入力データ	レコード番号「1050」入力データを 機械学習モデルに入力した際の出力結果
1051	レコード番号「1051」 入力データ	レコード番号「1051」入力データを 機械学習モデルに入力した際の出力結果
…	…	…

【図１４】

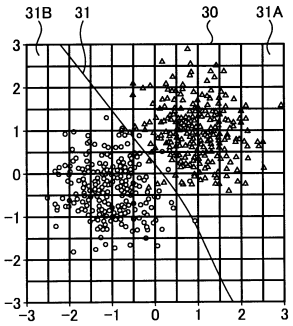
145

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

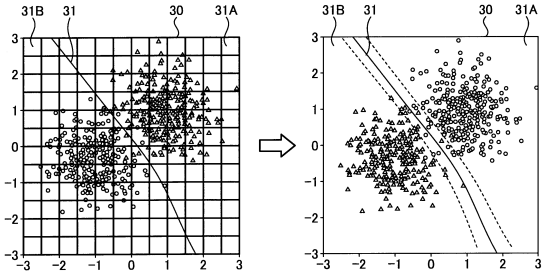
【図１５】



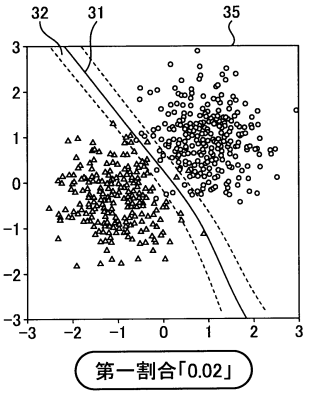
【図１６】



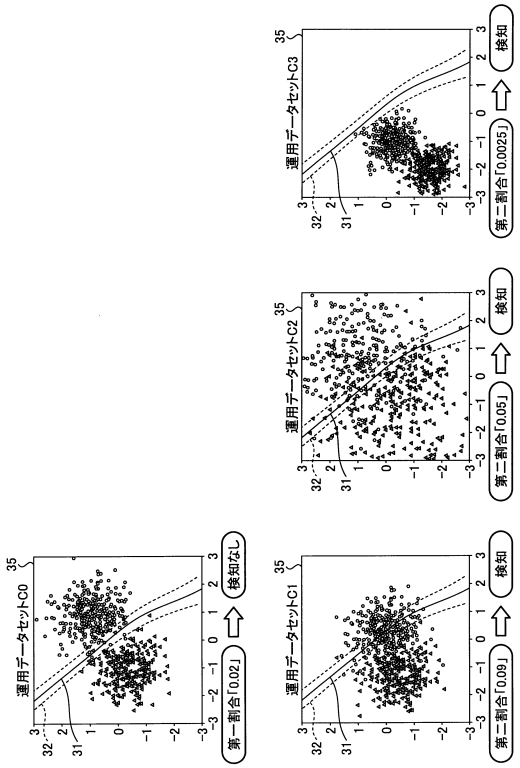
【図１７】



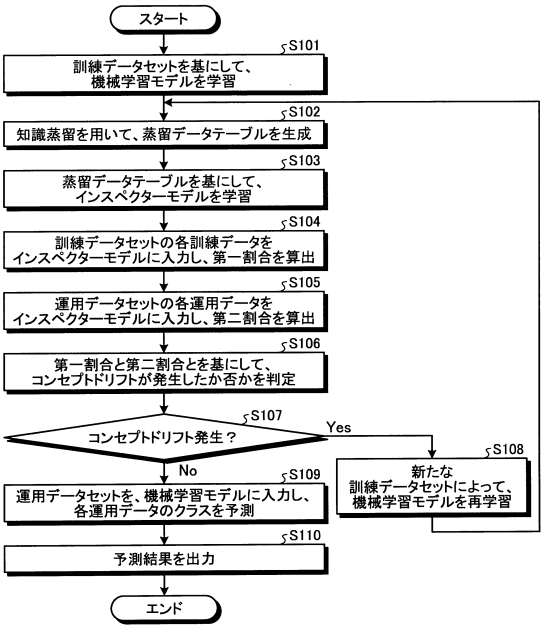
【図１８】



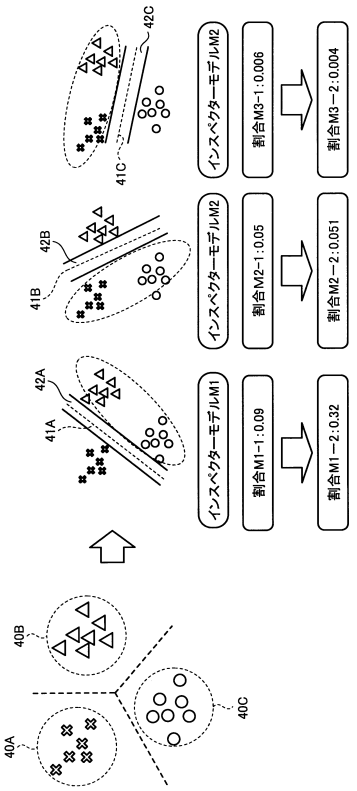
【図１９】



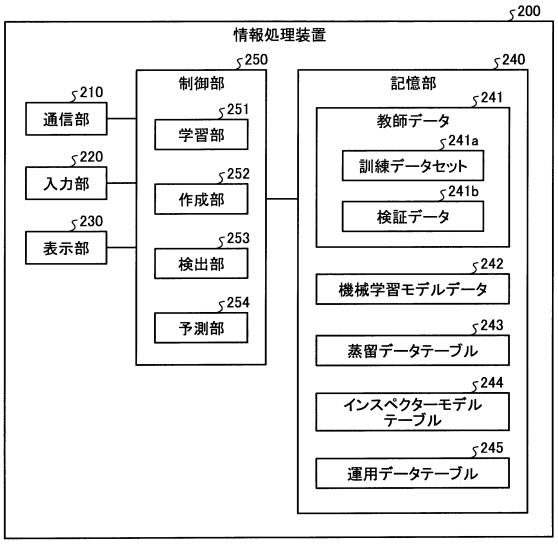
【図２０】



【図２１】



【図２２】

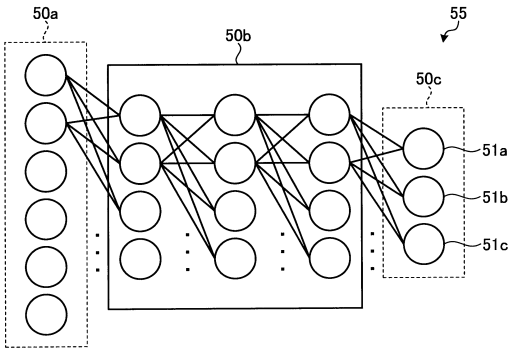


【図２３】

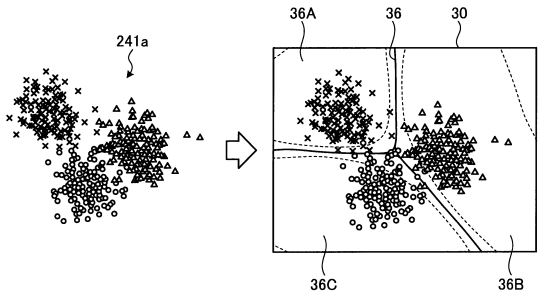
241a

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
・・・	・・・	・・・
1100	レコード番号「1100」 訓練データ	第3クラス
1101	レコード番号「1100」 訓練データ	第3クラス
・・・	・・・	・・・

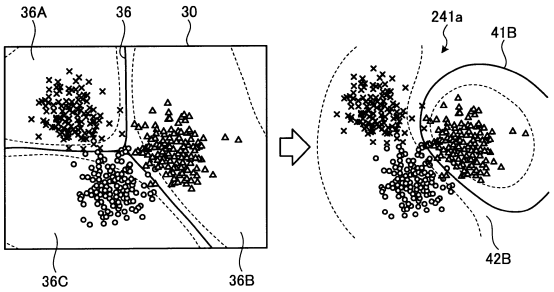
【図２４】



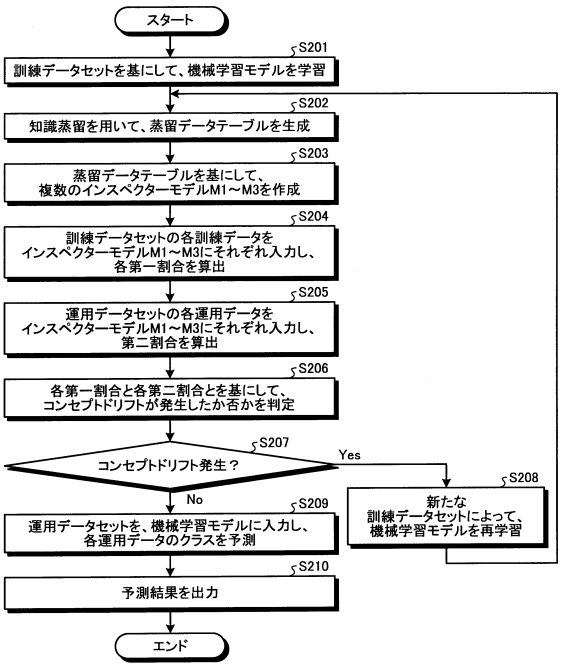
【図２５】



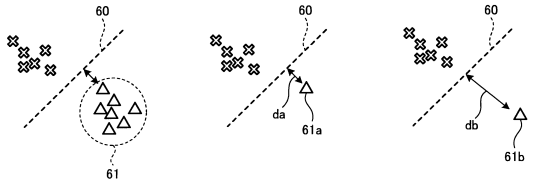
【図２６】



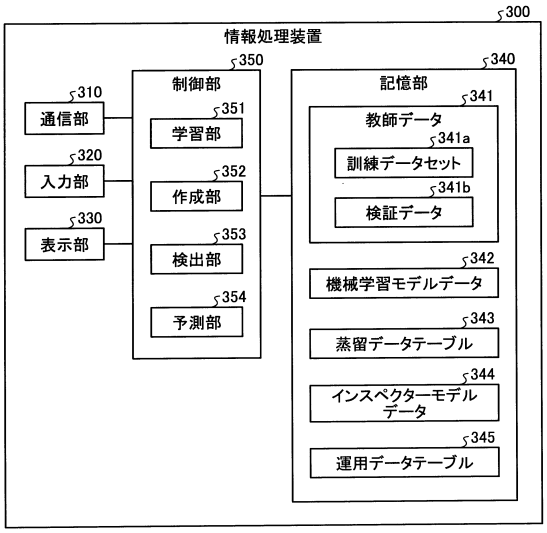
【図２７】



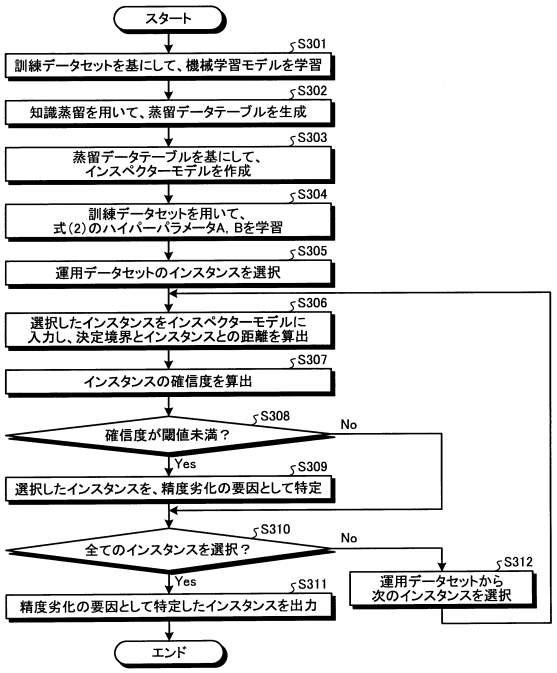
【図２８】



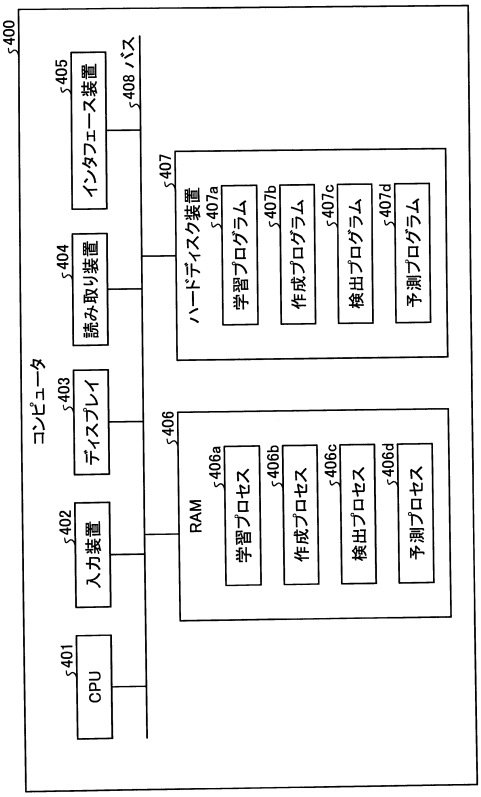
【図２９】



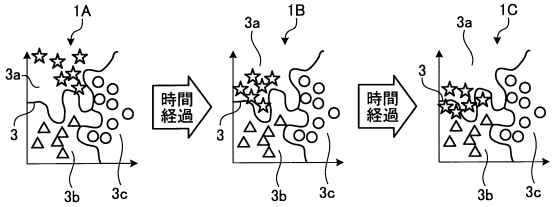
【図３０】



【図３１】



【図３２】



フロントページの続き

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ　３／００－９９／００