

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7306469号
(P7306469)

(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

F I

G 0 6 N 20/00 (2019.01)

G 0 6 N 20/00

請求項の数 5 (全36頁)

(21)出願番号	特願2021-553230(P2021-553230)	(73)特許権者	000005223
(86)(22)出願日	令和1年10月24日(2019.10.24)		富士通株式会社
(86)国際出願番号	PCT/JP2019/041691		神奈川県川崎市中原区上小田中4丁目1番1号
(87)国際公開番号	WO2021/079460	(74)代理人	110002147
(87)国際公開日	令和3年4月29日(2021.4.29)		弁理士法人酒井国際特許事務所
審査請求日	令和4年3月25日(2022.3.25)	(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、
複数の正解ラベル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、
前記運用モデルの出力結果を基にして、データの特徴空間を2つの適用領域に分割する決定境界から、入力された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、
訓練データを前記インスペクターモデルに入力して、前記決定境界から前記訓練データまでの第1の距離を算出し、
運用データを前記インスペクターモデルに入力して、前記決定境界から前記運用データまでの第2の距離を算出し、
前記第1の距離と、前記第2の距離との差が予め設定された閾値以上の場合に、前記運用データと前記訓練データとの間の特徴の違いに起因する前記運用モデルの出力結果の精度低下の可能性を検出する
処理を実行することを特徴とする検出方法。

【請求項2】

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て、前記決定境界から前記運用データまでの第2の距離をそれぞれ算出し、前記2の距離が予め設定された距離未満となる運用データを検出する処理を更に実行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】

前記第 2 の距離を、0 以上 1 未満の値をとる確信度に変換する処理を更に実行し、確信度が予め設定された値未満となる運用データを検出する処理を実行することを特徴とする請求項 1 または 2 に記載の検出方法。

【請求項 4】

コンピュータに、

複数の正解ラベル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、

前記運用モデルの出力結果を基にして、データの特徴空間を 2 つの適用領域に分割する決定境界から、入力された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、

訓練データを前記インスペクターモデルに入力して、前記決定境界から前記訓練データまでの第 1 の距離を算出し、

運用データを前記インスペクターモデルに入力して、前記決定境界から前記運用データまでの第 2 の距離を算出し、

前記第 1 の距離と、前記第 2 の距離との差が予め設定された閾値以上の場合に、前記運用データと前記訓練データとの間の特徴の違いに起因する前記運用モデルの出力結果の精度低下の可能性を検出する

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】

複数の正解ラベル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、

前記運用モデルの出力結果を基にして、データの特徴空間を 2 つの適用領域に分割する決定境界から、入力された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、

訓練データを前記インスペクターモデルに入力して、前記決定境界から前記訓練データまでの第 1 の距離を算出し、運用データを前記インスペクターモデルに入力して、前記決定境界から前記運用データまでの第 2 の距離を算出し、前記第 1 の距離と、前記第 2 の距離との差が予め設定された閾値以上の場合に、前記運用データと前記訓練データとの間の特徴の違いに起因する前記運用モデルの出力結果の精度低下の可能性を検出する検出部と

処理を実行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 3 2 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 3 2 において、分布 1 A は、システム運用初期の入力データの分布を示す。分布 1 B は、システム運用初期から T 1 時間経過した時点の入力データの分布を示す。分布 1 C は

10

20

30

40

50

、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。

10

【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。

20

【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、T²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T²統計量を算出する。T²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T²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T²統計量は、異常値データの割合に対応する。

30

【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.

40

【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、T²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。

50

【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。コンピュータが複数の正解ラベル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。コンピュータが、運用モデルの出力結果を基にして、データの特徴空間を2つの適用領域に分割する決定境界を学習すると共に、決定境界から入力された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。コンピュータが、訓練データをインスペクターモデルに入力して、決定境界から訓練データまでの第1の距離を算出する。コンピュータが、運用データをインスペクターモデルに入力して、決定境界から運用データまでの第2の距離を算出する。コンピュータが、第1の距離と、第2の距離との差が予め設定された閾値以上の場合に、運用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。

10

20

30

40

50

【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 t_1 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 t_1 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】

10

20

30

40

50

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 A₁ 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 A₁ が、分布 A₂ に変化する可能性がある。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 A₁ 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 11 A～11 C を作成している。そして、複数のインスペクターモデル 11 A～11 C 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 A₁ 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 5 a を監視し、危険領域 5 a 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に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 5 a 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 7 A の出力値を模倣するような、Student モデル 7 B 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 7 A および Student モデル 7 B を NN とするが、これに限定されるもので

10

20

30

40

50

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7A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） H_k 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 H_k 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】

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

【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 X_a 、 Y_a を特定する。情報処理装置は、サポートベクトル X_a および決定境界5のマージンと、サポートベクトル Y_a 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。

10

【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。

20

【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 (Random Forest)、NNをそれぞれ学習する。

30

【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ってくる。

40

【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】

50

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 20 A 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、22 A に示すものとなり、決定境界は、決定境界 23 A となる。

【0051】

情報処理装置は、分布 20 B 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、22 B に示すものとなり、決定境界は、決定境界 23 B となる。情報処理装置は、分布 20 C 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、22 C に示すものとなり、決定境界は、決定境界 23 C 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 23 A を基にした危険領域は、危険領域 24 A となる。第 2 インスペクターモデルの決定境界 23 B を基にした危険領域は、危険領域 24 B となる。第 3 インスペクターモデルの決定境界 23 C を基にした危険領域は、危険領域 24 C 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 141 a と、検証データ 141 b を有する。訓練データセット 141 a 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 141 b は、訓練データセット 141 a 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 141 b は、正解ラベルが付与される。たとえば、検証データ 141 b 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

10

20

30

40

50

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を通して、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「k SVM」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

10

20

30

40

50

）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。

10

【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。

20

【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】

30

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。

40

【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

50

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、インスぺ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、k SVM によって構築されたインスぺクターモデル 35 を作成する。作成部 152 は、作成したインスぺクターモデル 35 のデータ（インスぺクターモデルデータ 144）を、記憶部 140 に登録する。

10

【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、インスぺクターモデル（k SVM）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。

20

【0078】

図 10 の説明に戻る。検出部 153 は、インスぺクターモデル 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 141 a の各訓練データを、インスぺクターモデル 35 に入力する。検出部 153 が、訓練データをインスぺクターモデル 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 141 a 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。

30

【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、インスぺクターモデル 35 に入力する。検出部 153 が、運用データをインスぺクターモデル 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。

40

【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】

50

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。

10

【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。

20

【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。

30

【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。

40

【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】

50

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 141 a 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 141 a 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

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

定する。閾値は、予め設定されているものとする。図 2 1 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 $M1 \sim M3$ 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 2 2 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 2 2 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 2 3 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 2 3 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

10

20

30

40

50

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クラスのうちの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、k SVM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】

10

20

30

40

50

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 241 a を取得し、訓練データセット 241 a 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 241 a 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 241 a 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 241 b 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 241 a 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 241 a 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 36 A と、モデル適用領域 36 B と、モデル適用領域 36 C 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 36 A」と「モデル適用領域 36 B, 36 C」との決定境界（図 21 の決定境界 41 A 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 41 A 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 36 B」と「モデル適用領域 36 A, 36 C」との決定境界（図 21 の決定境界 41 B 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 41 B 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 36 C」と「モデル適用領域 36 A, 36 B」との

10

20

30

40

50

決定境界（図 2 1 の決定境界 4 1 C に相当）を学習した、インスペクターモデル M 3 を作成する。このインスペクターモデル M 3 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 4 1 C 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 2 6 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 2 6 では、一例として、インスペクターモデル M 2 の決定境界および危険領域 4 2 B を示す。インスペクターモデル M 1，M 3 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 2 5 2 は、蒸留データテーブル 2 4 3 を生成する処理、インスペクターモデルテーブル 2 4 4 を作成する処理を実行する。

【0140】

まず、作成部 2 5 2 が、蒸留データテーブル 2 4 3 を生成する処理について説明する。作成部 2 5 2 は、機械学習モデルデータ 2 4 2 を用いて、機械学習モデル 5 5 を実行し、特徴空間上の各データを、機械学習モデル 5 5 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 2 5 2 は、特徴空間をモデル適用領域 3 6 A と、モデル適用領域 3 6 B，モデル適用領域 3 6 C とに分類し、決定境界 3 6 を特定する。

【0141】

作成部 2 5 2 は、特徴空間 3 0 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 2 5 2 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 5 5 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 2 5 2 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 2 4 3 に登録する。作成部 2 5 2 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 2 4 3 を生成する。

【0142】

続いて、作成部 2 5 2 が、インスペクターモデルテーブル 2 4 4 を作成する処理について説明する。作成部 2 5 2 は、蒸留データテーブル 2 4 3 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、k S V M によって構築されたインスペクターモデル M 1 ~ M 3 を作成する。作成部 2 5 2 は、作成したインスペクターモデル M 1 ~ M 3 のデータを、インスペクターモデルテーブル 2 4 4 に登録する。

【0143】

作成部 2 5 2 が、「インスペクターモデル M 1」を作成する処理の一例について説明する。作成部 2 5 2 は、蒸留データテーブル 2 4 3 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 2 5 2 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 4 1 A 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 2 5 2 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 4 1 A 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 2 5 2 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 4 1 A を特定することで、インスペクターモデル M 1 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 4 1 A は直線となり、決定境界 4 1 A からの距離が m_{M1} となる領域を、危険領域 4 2 A に設定する。距離 m_{M1} は、決定境界 4 1 A 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 2 5 2 が、「インスペクターモデル M 2」を作成する処理の一例について説明する。作成部 2 5 2 は、蒸留データテーブル 2 4 3 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 2 5 2 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 4 1 B 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 2 5 2 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

10

20

30

40

50

クラスの入力データのうち、決定境界 4 1 B 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 2 5 2 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 4 1 B を特定することで、インスペクターモデル M 2 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 4 1 B は直線となり、決定境界 4 1 B からの距離が m_{M2} となる領域を、危険領域 4 2 B に設定する。距離 m_{M2} は、決定境界 4 1 B 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 2 5 2 が、「インスペクターモデル M 3」を作成する処理の一例について説明する。作成部 2 5 2 は、蒸留データテーブル 2 4 3 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 2 5 2 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 4 1 C 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 2 5 2 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 4 1 C 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 2 5 2 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 4 1 C を特定することで、インスペクターモデル M 3 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 4 1 C は直線となり、決定境界 4 1 C からの距離が m_{M3} となる領域を、危険領域 4 2 C に設定する。距離 m_{M3} は、決定境界 4 1 C 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 2 5 3 は、インスペクターモデル M 1 ~ M 3 を実行して、機械学習モデル 5 5 の精度劣化を検出する処理部である。また、検出部 2 5 3 は、機械学習モデル 5 5 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 2 5 3 は、インスペクターモデル M 1 ~ M 3 に訓練データセット 2 4 1 a をそれぞれ入力することで、各第一割合（割合 M 1 - 1、割合 M 2 - 1、割合 M 3 - 1）を算出する。

【0148】

検出部 2 5 3 は、訓練データを、インスペクターモデル M 1 に入力すると、特徴空間上の決定境界 4 1 A と訓練データとの距離が出力される。検出部 2 5 3 は、決定境界 4 1 A 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 4 2 A に含まれると判定する。検出部 2 5 3 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 4 2 A に含まれる訓練データの数を特定し、割合 M 1 - 1 を算出する。

【0149】

検出部 2 5 3 は、訓練データを、インスペクターモデル M 2 に入力すると、特徴空間上の決定境界 4 1 B と訓練データとの距離が出力される。検出部 2 5 3 は、決定境界 4 1 B 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 4 2 B に含まれると判定する。検出部 2 5 3 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 4 2 B に含まれる訓練データの数を特定し、割合 M 2 - 1 を算出する。

【0150】

検出部 2 5 3 は、訓練データを、インスペクターモデル M 3 に入力すると、特徴空間上の決定境界 4 1 C と訓練データとの距離が出力される。検出部 2 5 3 は、決定境界 4 1 C 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 4 2 C に含まれると判定する。検出部 2 5 3 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 4 2 C に含まれる訓練データの数を特定し、割合 M 3 - 1 を算出する。

【0151】

検出部 2 5 3 は、インスペクターモデル M 1 ~ M 3 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 M 1 - 2、割合 M 2 - 2、割合 M 3 - 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、精度劣化を検出した旨と、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、精度劣化を検出した旨と、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】

10

20

30

40

50

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 d_a である。インスタンス 61b と、決定境界 60 との距離は d_b である。距離 d_a は、距離 d_b 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。

10

【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。

20

【0175】

$$P(y=1|x) = 1 / (1 + \exp(Af(x) + B)) \cdots (2)$$

【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】

30

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。

40

【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

50

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、k 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。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】

10

20

30

40

50

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

10

20

30

40

50

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。

10

20

30

40

50

【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については、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】

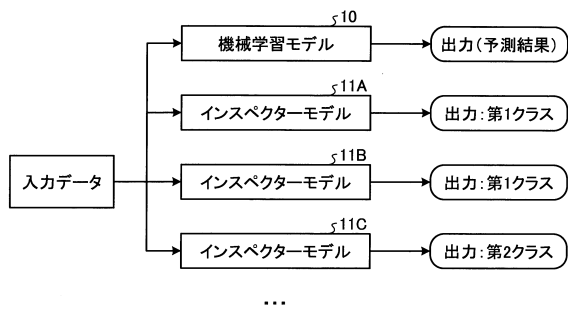
【0218】

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

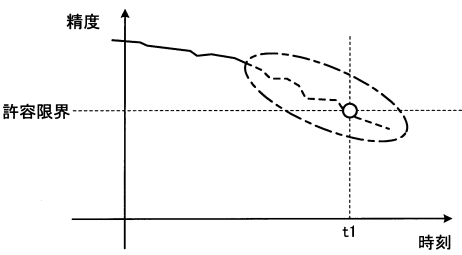
1 5 2 , 2 5 2 , 3 5 2 作成部
1 5 3 , 2 5 3 , 3 5 3 検出部
1 5 4 , 2 5 4 , 3 5 4 予測部
2 4 4 インспекターモデルテーブル

【図面】

【図 1】

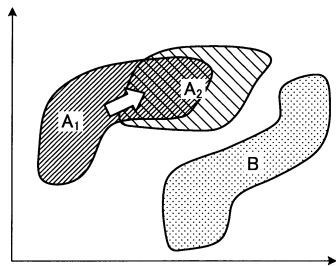


【図 2】

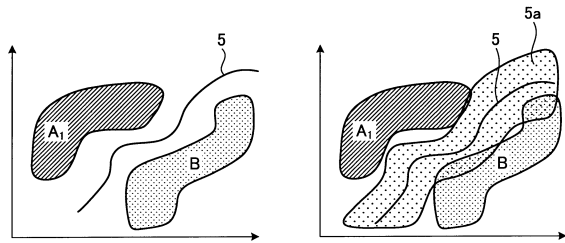


10

【図 3】



【図 4】



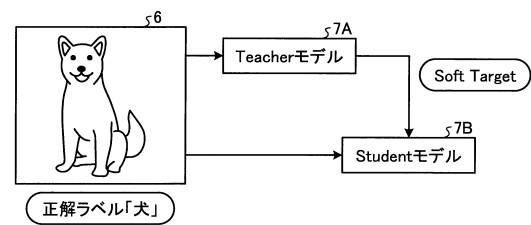
20

30

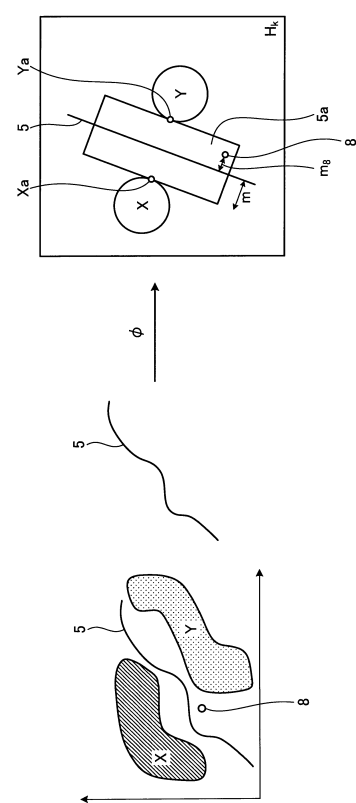
40

50

【図 5】



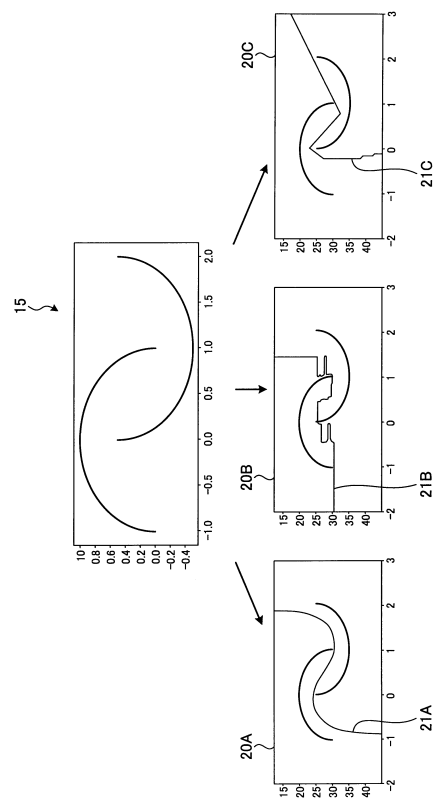
【図 6】



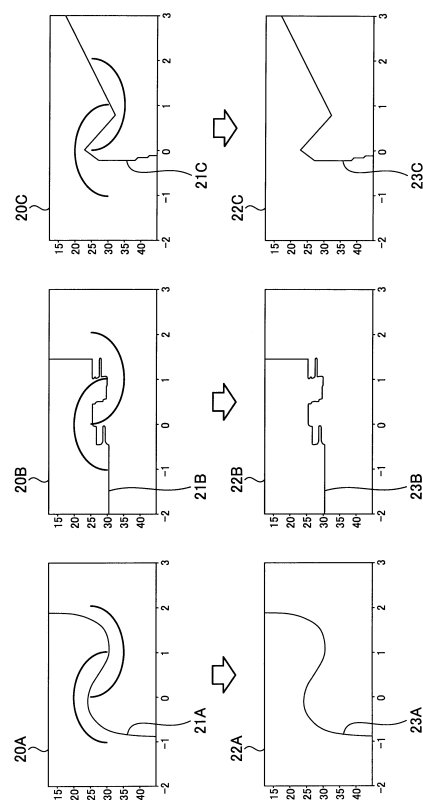
10

20

【図 7】



【図 8】

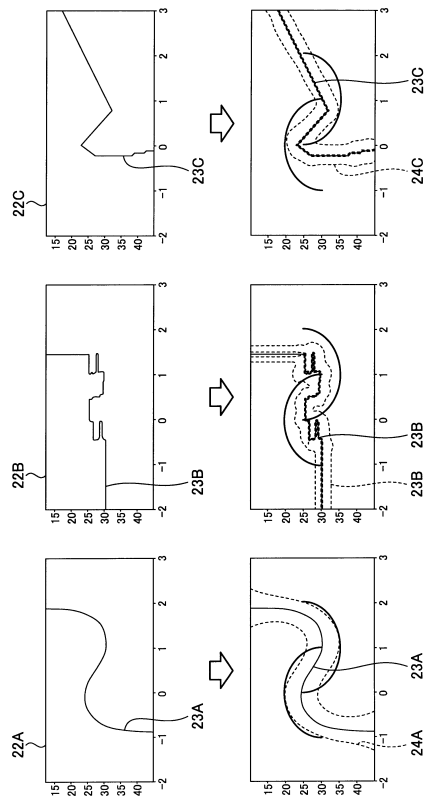


30

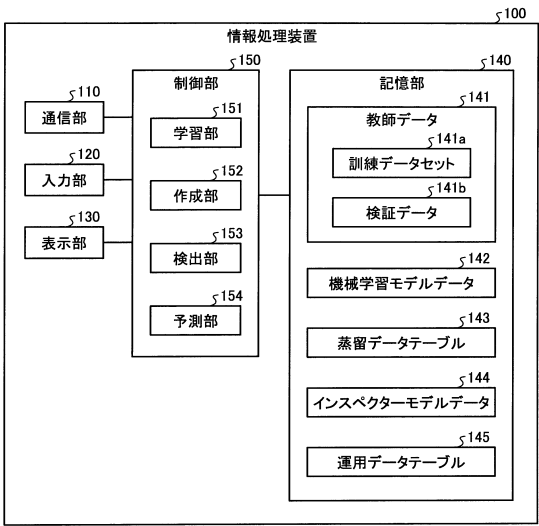
40

50

【図 9】



【図 10】



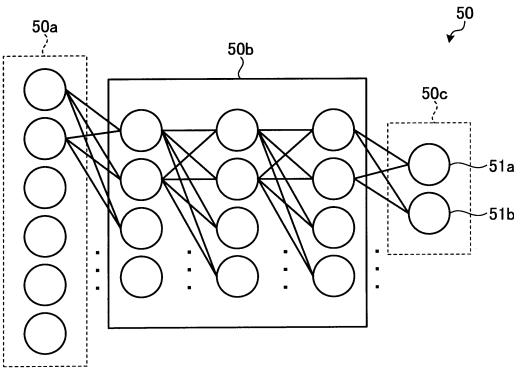
10

20

【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】



30

40

50

【図 1 3】

§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
...	...	...

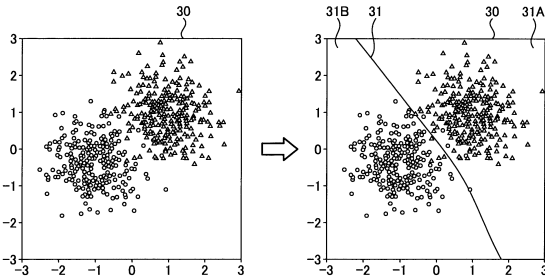
【図 1 4】

§145

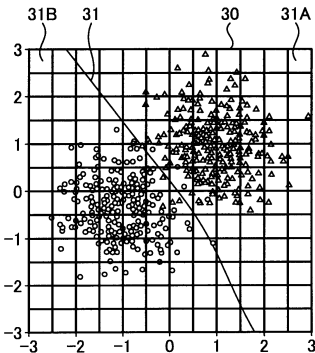
データ識別情報	運用データセット
C0	データ識別情報「C0」の運用データセット
C1	データ識別情報「C1」の運用データセット
C2	データ識別情報「C2」の運用データセット
C3	データ識別情報「C3」の運用データセット

10

【図 1 5】



【図 1 6】



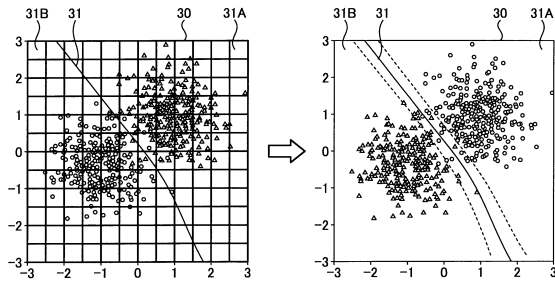
20

30

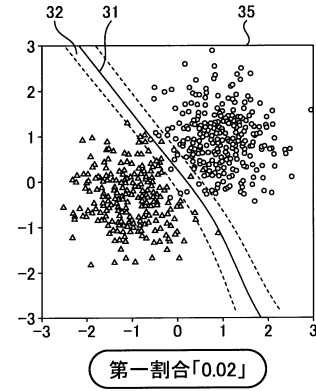
40

50

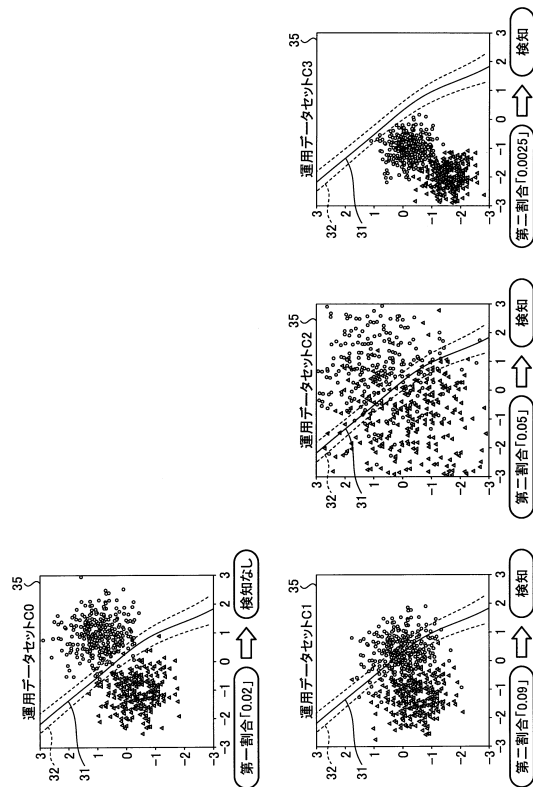
【図 17】



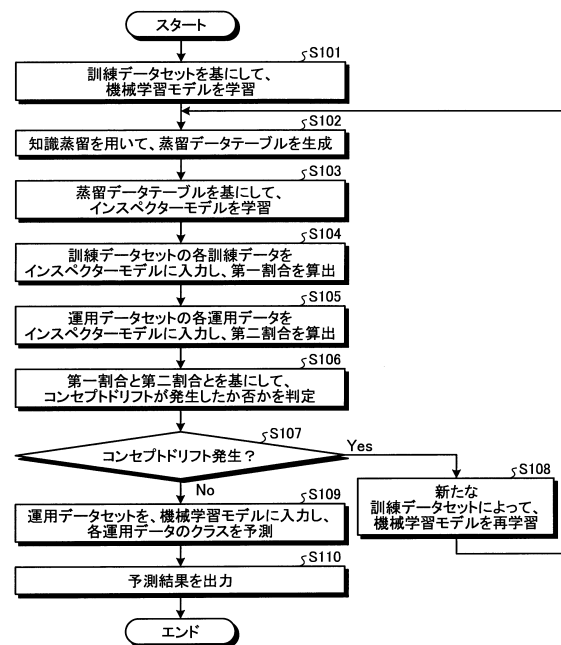
【図 18】



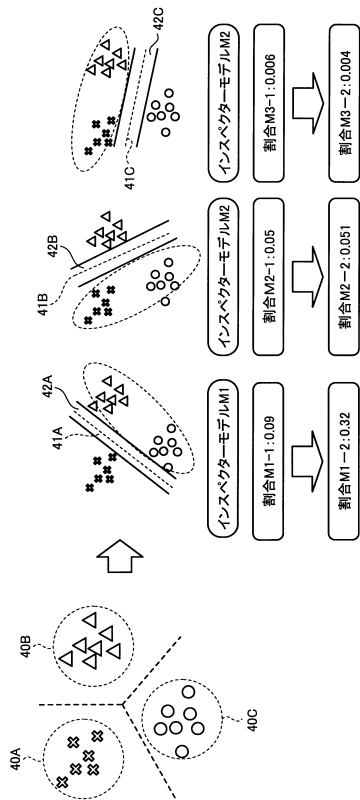
【図 19】



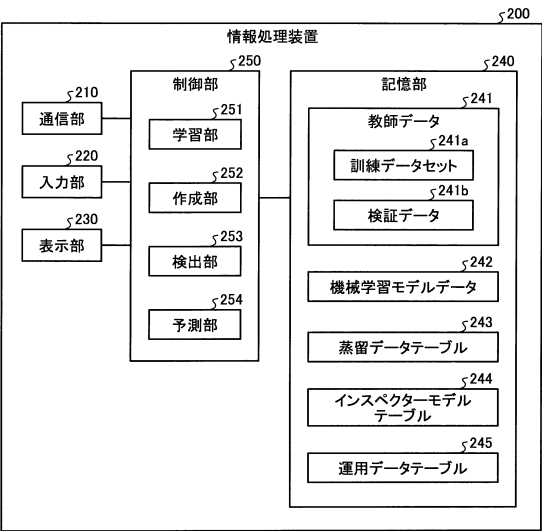
【図 20】



【図 2 1】



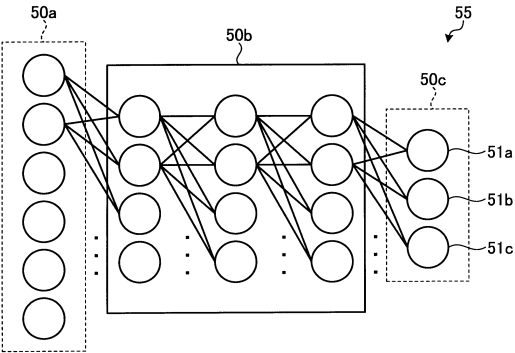
【図 2 2】



【図 2 3】

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 2 4】



10

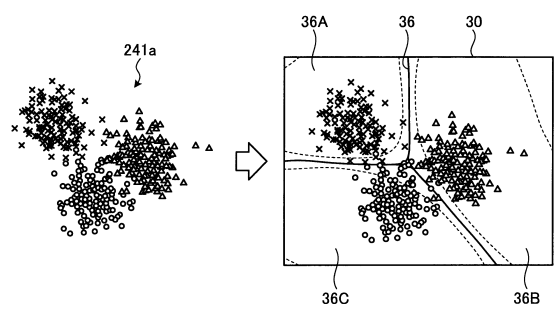
20

30

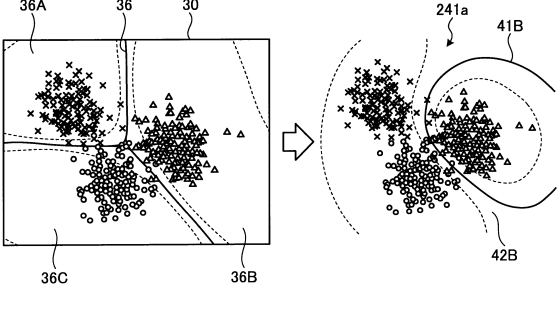
40

50

【図 2 5】

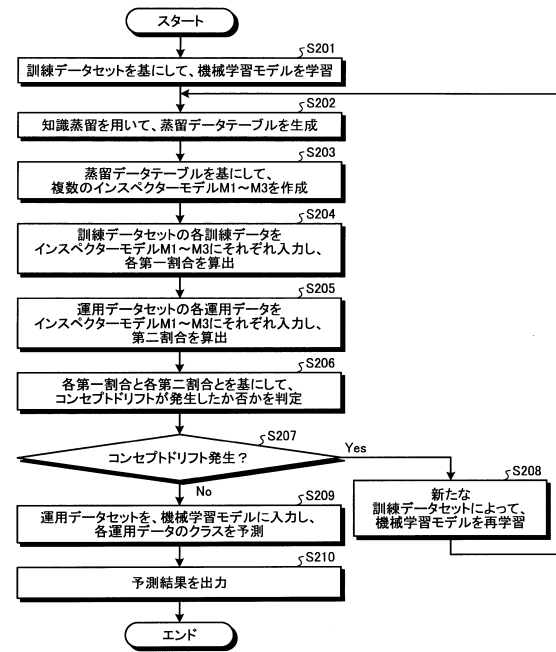


【図 2 6】



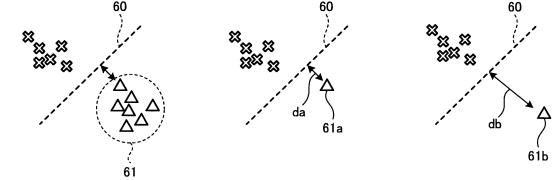
10

【図 2 7】



20

【図 2 8】

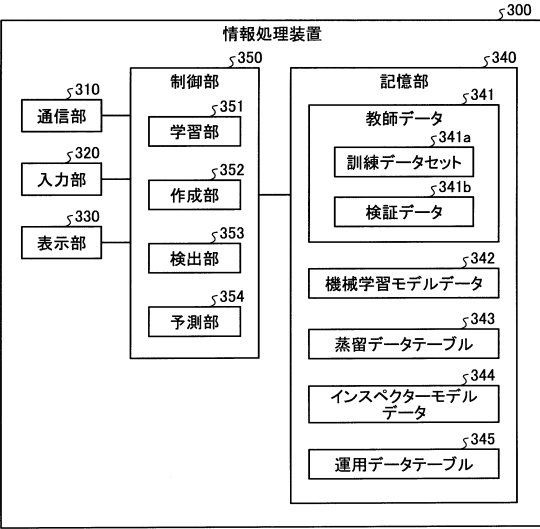


30

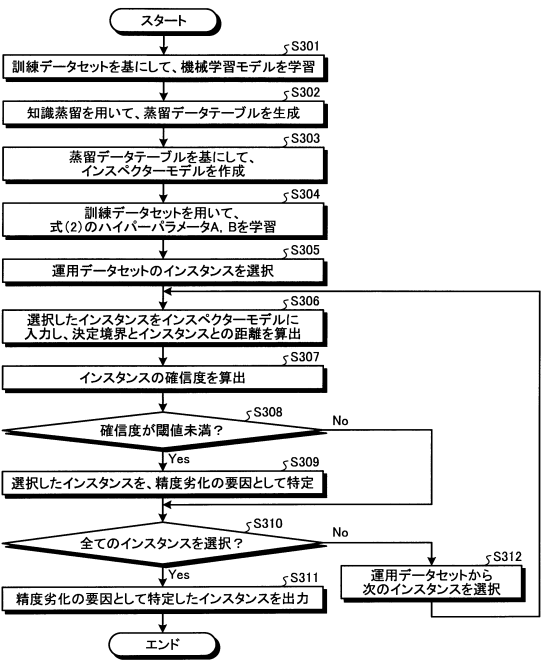
40

50

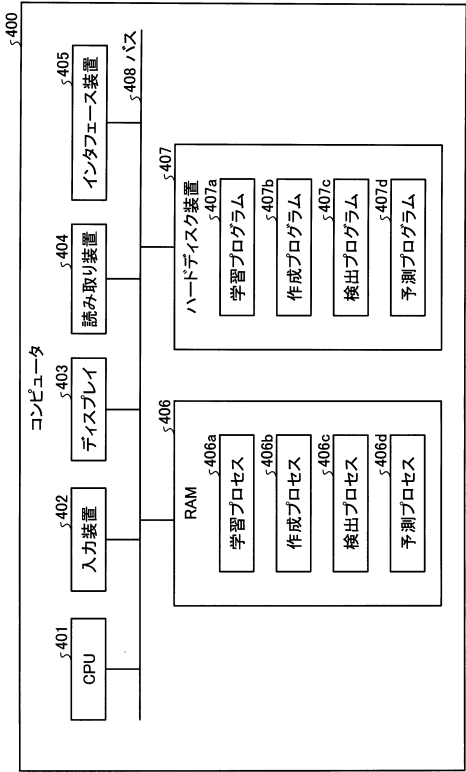
【図 29】



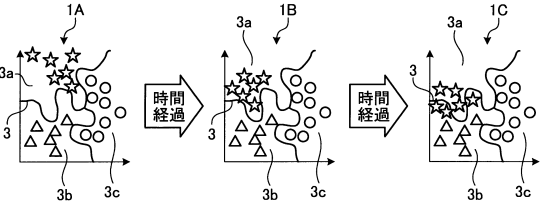
【図 30】



【図 31】



【図 32】



10

20

30

40

50

フロントページの続き

(58)調査した分野 (Int.Cl., D B名)

G 0 6 N 3 / 0 0 - 9 9 / 0 0