(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7231829号
(P7231829)

(45)発行日　令和5年3月2日(2023.3.2)

(24)登録日　令和5年2月21日(2023.2.21)

(51)国際特許分類
G06N 20/00 (2019.01)

FI
G06N 20/00

請求項の数　8　（全32頁）

(21)出願番号　特願2019-137027(P2019-137027)
(22)出願日　令和1年7月25日(2019.7.25)
(65)公開番号　特開2021-22051(P2021-22051A)
(43)公開日　令和3年2月18日(2021.2.18)
審査請求日　令和4年4月7日(2022.4.7)

(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　110002918
弁理士法人扶桑国際特許事務所
(72)発明者　小林 健一
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　松平 英

最終頁に続く

(54)【発明の名称】　機械学習プログラム、機械学習方法および機械学習装置

(57)【特許請求の範囲】

【請求項１】

コンピュータに、

データ集合から抽出された複数の第１の訓練データを用いて、機械学習により前記複数の第１の訓練データに対応する複数の第１のモデルを学習し、

前記データ集合から抽出された第１のテストデータに含まれる２以上のレコードそれぞれを前記複数の第１のモデルに入力することで、前記複数の第１のモデルと前記２以上のレコードとの組み合わせ毎に算出された予測誤差を示す誤差情報を生成し、

前記誤差情報に基づいて、テストデータのサイズとテストデータを用いて算出されるモデルの精度の測定値が有する分散との間の対応関係を判定し、

前記データ集合から抽出された第２の訓練データを用いて学習された第２のモデルの精度を、前記データ集合から抽出される第２のテストデータを用いて測定する場合に、前記対応関係に基づいて、前記第２のモデルに対して算出される精度の測定値の分散が所定条件を満たすように前記第２のテストデータのサイズを決定する、

処理を実行させる機械学習プログラム。

【請求項２】

前記対応関係は、テストデータのサイズの増加に応じて分散が下限に漸近するように減少する非線形関係であり、前記第２のテストデータのサイズは、サイズの所定増加量に対する分散の減少度を示す効率性指標に基づいて決定される、

請求項１記載の機械学習プログラム。

【請求項３】
前記所定条件は、前記効率性指標の値が閾値以上であることであり、前記第２のテストデータのサイズは、前記所定条件を満たす範囲で最大のサイズに決定される、
請求項２記載の機械学習プログラム。
【請求項４】
前記対応関係の判定では、前記２以上のレコードそれぞれについて前記複数の第１のモデルに対して算出された予測誤差を平均化した予測バイアスを算出し、前記２以上のレコードの前記予測バイアスを合成して、前記対応関係を表すパラメータの値を決定する、
請求項１記載の機械学習プログラム。
【請求項５】
前記対応関係の判定では、訓練データのサイズに依存しない第１のパラメータと訓練データのサイズに依存する第２のパラメータとテストデータのサイズを示す第３のパラメータとを用いて分散を算出する分散関数に対して、前記第１のパラメータの値を推定し、
前記第２のテストデータのサイズの決定では、前記第２のモデルの学習結果に基づいて前記第２のパラメータの値を推定し、前記第３のパラメータの値を変動させることで、分散が前記所定条件を満たすテストデータのサイズを探索する、
請求項１記載の機械学習プログラム。
【請求項６】
前記第２のテストデータのサイズの決定では、前記第１のテストデータを前記第２のモデルに入力して算出される予測誤差に基づいて前記第２のパラメータの値を仮選択し、前記仮選択した第２のパラメータの値を用いてテストデータのサイズを仮選択し、前記データ集合から抽出された前記仮選択したサイズのテストデータを前記第２のモデルに入力して算出される予測誤差に基づいて前記第２のパラメータの値を決定する、
請求項５記載の機械学習プログラム。
【請求項７】
コンピュータが、
データ集合から抽出された複数の第１の訓練データを用いて、機械学習により前記複数の第１の訓練データに対応する複数の第１のモデルを学習し、
前記データ集合から抽出された第１のテストデータに含まれる２以上のレコードそれぞれを前記複数の第１のモデルに入力することで、前記複数の第１のモデルと前記２以上のレコードとの組み合わせ毎に算出された予測誤差を示す誤差情報を生成し、
前記誤差情報に基づいて、テストデータのサイズとテストデータを用いて算出されるモデルの精度の測定値が有する分散との間の対応関係を判定し、
前記データ集合から抽出された第２の訓練データを用いて学習された第２のモデルの精度を、前記データ集合から抽出される第２のテストデータを用いて測定する場合に、前記対応関係に基づいて、前記第２のモデルに対して算出される精度の測定値の分散が所定条件を満たすように前記第２のテストデータのサイズを決定する、
機械学習方法。
【請求項８】
データ集合を記憶する記憶部と、
前記データ集合から抽出された複数の第１の訓練データを用いて、機械学習により前記複数の第１の訓練データに対応する複数の第１のモデルを学習し、前記データ集合から抽出された第１のテストデータに含まれる２以上のレコードそれぞれを前記複数の第１のモデルに入力することで、前記複数の第１のモデルと前記２以上のレコードとの組み合わせ毎に算出された予測誤差を示す誤差情報を生成し、前記誤差情報に基づいて、テストデータのサイズとテストデータを用いて算出されるモデルの精度の測定値が有する分散との間の対応関係を判定し、前記データ集合から抽出された第２の訓練データを用いて学習された第２のモデルの精度を、前記データ集合から抽出される第２のテストデータを用いて測定する場合に、前記対応関係に基づいて、前記第２のモデルに対して算出される精度の測定値の分散が所定条件を満たすように前記第２のテストデータのサイズを決定する処理部

と、
を有する機械学習装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は機械学習プログラム、機械学習方法および機械学習装置に関する。
【背景技術】
【０００２】
コンピュータを利用したデータ分析の１つとして、機械学習が行われることがある。機械学習では、幾つかの既知の事例を示す訓練データをコンピュータに入力する。コンピュータは、訓練データを分析して、要因（説明変数や独立変数と言うことがある）と結果（目的変数や従属変数と言うことがある）との間の関係を一般化したモデルを学習する。学習されたモデルを用いることで、未知の事例についての結果を予測することができる。
【０００３】
機械学習では、学習されるモデルの正確さ、すなわち、未知の事例の結果を正確に予測する能力（モデルの精度、予測性能、性能などと言うことがある）が問題となる。モデルの精度は、分析対象とする事象の性質、モデルの学習に使用する訓練データのサイズ、機械学習アルゴリズムなど様々な側面に依存する。精度が不十分なモデルが生成された場合、訓練データのサイズを増加させる、機械学習アルゴリズムを変更するなどの調整を行って、機械学習を再実行することもある。そこで、データ母集合から訓練データとは異なる既知の事例を示すテストデータを抽出し、訓練データを用いて学習されたモデルに対してテストデータを入力することで、モデルの精度を評価することが多い。
【０００４】
なお、同一のデータ母集合から、異なる分割方法によって訓練データとテストデータのペアを複数通り生成し、ペア毎に訓練データを用いたモデルの学習とテストデータを用いた精度の評価を行い、精度の平均を算出する汎化能力評価方法が提案されている。また、データベースから訓練データを抽出して回帰分析を行い、回帰モデルの精度を評価し、精度が不十分である場合には訓練データを追加して回帰分析を再度行う結果予測装置が提案されている。また、教師ラベルが付されたサンプルのデータ母集合から、訓練データと類似するテストデータを抽出し、訓練データを用いて学習された分類モデルの精度を、訓練データと類似するテストデータを用いて評価する情報処理システムが提案されている。
【先行技術文献】
【特許文献】
【０００５】
【文献】特開平９－５４７６４号公報
特開２０１４－１３５６０号公報
国際公開第２０１７／１８３５４８号
【発明の概要】
【発明が解決しようとする課題】
【０００６】
しかし、モデルの精度を評価するにあたり、テストデータのサイズをどの様に決定すればよいかが問題となる。テストデータが少な過ぎると、テストデータとして選択されるサンプルの偶然性の影響を強く受けて、算出される精度が不正確になり信頼性が低下する。一方、テストデータが多過ぎると、精度の評価に長時間かかることになり非効率である。この点、従来の機械学習では、訓練データのサイズの２分の１から４分の１程度をテストデータのサイズとするなど、経験則に基づいてサイズを決定していた。そのため、テストデータを用いたモデルの精度の評価について改善の余地があった。
【０００７】
１つの側面では、本発明は、機械学習のテストデータのサイズを適切に決定できる機械学習プログラム、機械学習方法および機械学習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】
【０００８】
１つの態様では、コンピュータに以下の処理を実行させる機械学習プログラムが提供される。データ集合から抽出された複数の第１の訓練データを用いて、機械学習により複数の第１の訓練データに対応する複数の第１のモデルを学習する。データ集合から抽出された第１のテストデータに含まれる２以上のレコードそれぞれを複数の第１のモデルに入力することで、複数の第１のモデルと２以上のレコードとの組み合わせ毎に算出された予測誤差を示す誤差情報を生成する。誤差情報に基づいて、テストデータのサイズとテストデータを用いて算出されるモデルの精度の測定値が有する分散との間の対応関係を判定する。データ集合から抽出された第２の訓練データを用いて学習された第２のモデルの精度を、データ集合から抽出される第２のテストデータを用いて測定する場合に、対応関係に基づいて、第２のモデルに対して算出される精度の測定値の分散が所定条件を満たすように第２のテストデータのサイズを決定する。
【０００９】
また、１つの態様では、コンピュータが実行する機械学習方法が提供される。また、１つの態様では、記憶部と処理部とを有する機械学習装置が提供される。
【発明の効果】
【００１０】
１つの側面では、機械学習のテストデータのサイズが適切に決定される。
【図面の簡単な説明】
【００１１】
【図１】第１の実施の形態の機械学習装置の例を説明する図である。
【図２】第２の実施の形態の機械学習装置のハードウェア例を示す図である。
【図３】訓練データサイズと予測性能の関係例を示すグラフである。
【図４】予測性能の測定値の分散例を示すグラフである。
【図５】予測性能の期待ロスおよび期待バイアスの例を示すグラフである。
【図６】機械学習装置の機能例を示すブロック図である。
【図７】誤差プロファイルテーブルの例を示す図である。
【図８】分散関数テーブルの例を示す図である。
【図９】機械学習の手順例を示すフローチャートである。
【図１０】機械学習の手順例を示すフローチャート（続き）である。
【発明を実施するための形態】
【００１２】
以下、本実施の形態を図面を参照して説明する。
［第１の実施の形態］
第１の実施の形態を説明する。
【００１３】
図１は、第１の実施の形態の機械学習装置の例を説明する図である。
第１の実施の形態の機械学習装置１０は、訓練データを用いて機械学習によりモデルを生成し、テストデータを用いてモデルの精度を測定する。機械学習装置１０を、情報処理装置やコンピュータと言うこともできる。機械学習装置１０は、ユーザが操作するクライアント装置でもよいし、他の装置からアクセスされるサーバ装置でもよい。
【００１４】
機械学習装置１０は、記憶部１１および処理部１２を有する。記憶部１１は、ＲＡＭ（Random Access Memory）などの揮発性半導体メモリでもよいし、ＨＤＤ（Hard Disk Drive）やフラッシュメモリなどの不揮発性ストレージでもよい。処理部１２は、例えば、ＣＰＵ（Central Processing Unit）、ＧＰＵ（Graphics Processing Unit）、ＤＳＰ（Digital Signal Processor）などのプロセッサである。ただし、処理部１２は、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field Programmable Gate Array）などの特定用途の電子回路を含んでもよい。プロセッサは、ＲＡＭなどのメ

モリ（記憶部１１でもよい）に記憶されたプログラムを実行する。複数のプロセッサの集合を「マルチプロセッサ」または単に「プロセッサ」と言うこともある。

【００１５】

記憶部１１は、データ集合１３を記憶する。データ集合１３は、既知の事例を示す複数のレコードを含む。レコードを、サンプルや行やデータブロックと言うこともできる。各レコードは、１以上の説明変数の値と１つの目的変数の値とを含む。説明変数を、カラムと言うこともできる。目的変数の値は、ユーザから与えられる正解であり、教師ラベルと言うこともできる。データ集合１３は、１００万レコード以上の多数のレコードを含んでもよく、ビッグデータと言われる大規模データであってもよい。

【００１６】

処理部１２は、データ集合１３から、訓練データ１４ａ，１４ｂ，１４ｃを含む複数セットの訓練データ（第１の訓練データ）を抽出する。ここで抽出する訓練データのセット数は少数でよく、各訓練データのサイズは小さくてよい。例えば、訓練データのセット数を１０セット程度とし、各訓練データのサイズを１万レコード程度とする。各訓練データのサイズは、後述する訓練データ１８のサイズの１００分の１程度でよい。処理部１２は、同一のレコードが異なる訓練データに含まれないようにデータ集合１３からレコードを抽出してもよいし、同一のレコードが異なる訓練データに含まれることを許容してもよい。処理部１２は、データ集合１３からランダムにレコードを抽出してもよい。

【００１７】

また、処理部１２は、データ集合１３からテストデータ１５（第１のテストデータ）を抽出する。テストデータ１５のサイズは、訓練データ１４ａ，１４ｂ，１４ｃのサイズより小さくてよい。テストデータ１５のサイズは、訓練データ１４ａ，１４ｂ，１４ｃのサイズの２分の１から４分の１程度でもよく、例えば、５０００レコード程度とする。処理部１２は、テストデータ１５に属するレコードを、訓練データ１４ａ，１４ｂ，１４ｃと重複しないようにデータ集合１３から抽出することが好ましい。

【００１８】

処理部１２は、訓練データ１４ａ，１４ｂ，１４ｃを含む複数セットの訓練データを用いて、機械学習によりそれら複数セットの訓練データに対応する複数のモデルを学習する。訓練データ１４ａから１つのモデルが学習され、それと独立に訓練データ１４ｂから１つのモデルが学習され、それと独立に訓練データ１４ｃから１つのモデルが学習される。

【００１９】

複数のモデルの学習には、同一の機械学習アルゴリズムが使用される。使用する機械学習アルゴリズムは、ユーザにより指定されてもよい。機械学習アルゴリズムとして、回帰分析、サポートベクタマシン、ランダムフォレストなどが挙げられる。モデルは、説明変数と目的変数との間の関係を示し、通常、１以上の説明変数と１以上の係数と１つの目的変数とを含む。係数は、機械学習を通じて訓練データに基づいて決定される。

【００２０】

次に、処理部１２は、テストデータ１５および学習した複数のモデルを用いて、誤差情報１６を生成する。誤差情報１６を、誤差プロファイルと言うこともできる。誤差情報１６は、複数セットの訓練データに対応する複数のモデルとテストデータ１５に含まれる２以上のレコードとの組み合わせ毎に算出された予測誤差を示す。

【００２１】

このとき、処理部１２は、ある訓練データから学習された１つのモデルに、テストデータ１５に含まれる１つのレコードを入力することで、当該１つのモデルと当該１つのレコードの組に対応する１つの予測誤差を算出する。例えば、処理部１２は、テストデータ１５のレコードに含まれる説明変数の値をモデルの説明変数に代入する。処理部１２は、モデルによって算出される目的変数の値である予測値と、テストデータ１５のレコードに含まれる目的変数の値である正解値とを比較し、両者の差を予測誤差として算出する。

【００２２】

次に、処理部１２は、誤差情報１６に基づいて対応関係１７を判定する。対応関係１７

は、テストデータのサイズと、テストデータを用いて算出されるモデルの精度の測定値が有するばらつきの程度である分散との間の対応関係を示す。モデルの精度は、未知の事例の結果を正確に予測する能力であり、予測性能や性能と言うこともできる。モデルの精度の指標として、正答率（Accuracy）、適合率（Precision）、平均二乗誤差（ＭＳＥ）、二乗平均平方根誤差（ＲＭＳＥ）などが挙げられる。

【００２３】

対応関係１７は、例えば、テストデータのサイズの増加に応じて分散が下限に漸近するように減少する非線形関係である。一般に、データ集合１３からのテストデータの抽出には、レコードの選択の偶然性がある。このため、テストデータのサイズが小さいと、レコードの選択の偶然性の影響を強く受けて、精度の測定値が真の値からずれるリスクが高くなる。テストデータのサイズを大きくすることで、分散を小さくすることができる。ただし、データ集合１３からの訓練データの抽出にも、レコードの選択の偶然性がある。テストデータのサイズの増加だけでは、精度の測定値の分散は０にならない。

【００２４】

対応関係１７は、機械学習に使用するデータ集合１３や機械学習アルゴリズムに依存し得る。そこで、処理部１２は、誤差情報１６に基づいて対応関係１７を判定する。例えば、処理部１２は、誤差情報１６が示す予測誤差のうち、テストデータ１５のレコードが同一でモデルが異なる予測誤差を平均化することで、テストデータ１５のレコード毎に予測バイアスを算出する。処理部１２は、テストデータ１５の２以上のレコードの予測バイアスを合成して、対応関係１７を規定するパラメータの値を決定する。

【００２５】

対応関係１７は、訓練データのサイズに依存しない第１のパラメータと、訓練データのサイズに依存する第２のパラメータと、テストデータのサイズを示す第３のパラメータとから分散を算出する分散関数であってもよい。この場合、処理部１２は、誤差情報１６を用いて第１のパラメータの値を推定してもよい。これにより、分散関数は、変数として第２のパラメータと第３のパラメータをもつ関数になる。

【００２６】

次に、処理部１２は、データ集合１３から訓練データ１８（第２の訓練データ）を抽出する。訓練データ１８のサイズは、訓練データ１４ａ，１４ｂ，１４ｃより十分に大きくてもよく、ユーザから指定されてもよい。例えば、訓練データ１８のサイズを１００万レコード程度とする。処理部１２は、訓練データ１８を用いてモデルを学習する。

【００２７】

モデルが学習されると、処理部１２は、データ集合１３からテストデータ１９（第２のテストデータ）を抽出する。処理部１２は、テストデータ１９に属するレコードを、訓練データ１８と重複しないようにデータ集合１３から抽出することが好ましい。処理部１２は、訓練データ１８から学習されたモデルの精度を、テストデータ１９を用いて測定する。例えば、処理部１２は、テストデータ１９のレコードに含まれる説明変数の値をモデルの説明変数に代入し、モデルによって算出される目的変数の予測値とテストデータ１９のレコードに含まれる目的変数の正解値とを比較して、精度を測定する。

【００２８】

このとき、処理部１２は、対応関係１７に基づいて、モデルの精度の測定値の分散が所定条件を満たすように、テストデータ１９のサイズを決定する。例えば、処理部１２は、対応関係１７において、サイズの所定増加量に対する分散の減少度を示す効率性指標を算出し、効率性指標に基づいてテストデータ１９のサイズを決定する。対応関係１７が、テストデータのサイズの増加に応じて分散が下限に漸近する非線形関係である場合、効率性指標の値は、テストデータのサイズの増加に応じて減少する。テストデータ１９のサイズは、効率性指標の値が閾値以上である範囲で最大のサイズとしてもよい。

【００２９】

また、例えば、処理部１２は、訓練データ１８を用いたモデルの学習結果に基づいて、分散関数に含まれる訓練データのサイズに依存する第２のパラメータの値を決定する。そ

して、処理部１２は、決定された上記の第１のパラメータの値および第２のパラメータの値のもとで、テストデータのサイズを示す第３のパラメータの値を変動させることで、分散が所定条件を満たすテストデータのサイズを探索する。

【００３０】

なお、誤差情報１６の生成および対応関係１７の判定は、訓練データ１８を用いた機械学習の前に行ってもよいし後に行ってもよい。処理部１２は、訓練データ１８を用いて学習されたモデルと、テストデータ１９を用いて測定された精度を出力する。処理部１２は、学習されたモデルと測定された精度を、記憶装置に保存してもよいし、表示装置に表示してもよいし、他の情報処理装置に送信してもよい。

【００３１】

第１の実施の形態の機械学習装置１０によれば、小さいサイズの訓練データ１４ａ，１４ｂ，１４ｃを用いて複数のモデルが学習される。小さいサイズのテストデータ１５を用いて、それら複数のモデルとテストデータ１５の２以上のレコードとの組み合わせ毎に算出された予測誤差を示す誤差情報１６が生成される。誤差情報１６に基づいて、テストデータのサイズとモデルの精度の測定値が有する分散との間の対応関係１７が判定される。そして、訓練データ１８を用いて学習されたモデルの精度を、テストデータ１９を用いて測定するにあたり、対応関係１７に基づいて、精度の測定値の分散が所定条件を満たすようにテストデータ１９のサイズが決定される。

【００３２】

これにより、テストデータ１９のサイズがモデル精度の測定値の分散に与える影響を考慮して、テストデータ１９のサイズを適切に決定することができる。よって、テストデータ１９のサイズが小さ過ぎることによる測定値の信頼性の低下を抑制できる。また、テストデータ１９のサイズが大き過ぎることによる処理時間の増大を抑制できる。このため、機械学習により学習されたモデルの精度を、高信頼かつ短時間で測定することができ、モデルの精度の測定を効率化できる。特に、テストデータのサイズを訓練データのサイズの２分の１から４分の１程度とする経験則と比べて、テストデータのサイズを削減できる。

【００３３】

［第２の実施の形態］

次に、第２の実施の形態を説明する。

図２は、第２の実施の形態の機械学習装置のハードウェア例を示す図である。

【００３４】

機械学習装置１００は、ＣＰＵ１０１、ＲＡＭ１０２、ＨＤＤ１０３、画像インタフェース１０４、入力インタフェース１０５、媒体リーダ１０６および通信インタフェース１０７を有する。機械学習装置１００が有するこれらのユニットは、機械学習装置１００の内部でバスに接続されている。機械学習装置１００は、第１の実施の形態の機械学習装置１０に対応する。ＣＰＵ１０１は、第１の実施の形態の処理部１２に対応する。ＲＡＭ１０２またはＨＤＤ１０３は、第１の実施の形態の記憶部１１に対応する。

【００３５】

ＣＰＵ１０１は、プログラムの命令を実行するプロセッサである。ＣＰＵ１０１は、ＨＤＤ１０３に記憶されたプログラムやデータの少なくとも一部をＲＡＭ１０２にロードし、プログラムを実行する。ＣＰＵ１０１は複数のプロセッサコアを備えてもよく、機械学習装置１００は複数のプロセッサを備えてもよい。複数のプロセッサの集合を「マルチプロセッサ」または単に「プロセッサ」と言うことがある。

【００３６】

ＲＡＭ１０２は、ＣＰＵ１０１が実行するプログラムやＣＰＵ１０１が演算に使用するデータを一時的に記憶する揮発性半導体メモリである。機械学習装置１００は、ＲＡＭ以外の種類のメモリを備えてもよく、複数のメモリを備えてもよい。

【００３７】

ＨＤＤ１０３は、ＯＳ（Operating System）やミドルウェアやアプリケーションソフトウェアなどのソフトウェアのプログラム、および、データを記憶する不揮発性ストレー

ジである。機械学習装置１００は、フラッシュメモリやＳＳＤ（Solid State Drive）など他の種類のストレージを備えてもよく、複数のストレージを備えてもよい。

【００３８】

画像インタフェース１０４は、ＣＰＵ１０１からの命令に従って、機械学習装置１００に接続された表示装置１１１に画像を出力する。表示装置１１１として、ＣＲＴ（Cathode Ray Tube）ディスプレイ、液晶ディスプレイ（ＬＣＤ：Liquid Crystal Display）、有機ＥＬ（ＯＥＬ：Organic Electro-Luminescence）ディスプレイ、プロジェクタなど、任意の種類の表示装置を使用することができる。機械学習装置１００に、プリンタなど表示装置１１１以外の出力デバイスが接続されてもよい。

【００３９】

入力インタフェース１０５は、機械学習装置１００に接続された入力デバイス１１２から入力信号を受け付ける。入力デバイス１１２として、マウス、タッチパネル、タッチパッド、キーボードなど、任意の種類の入力デバイスを使用することができる。機械学習装置１００に複数種類の入力デバイスが接続されてもよい。

【００４０】

媒体リーダ１０６は、記録媒体１１３に記録されたプログラムやデータを読み取る読み取り装置である。記録媒体１１３として、フレキシブルディスク（ＦＤ：Flexible Disk）やＨＤＤなどの磁気ディスク、ＣＤ（Compact Disc）やＤＶＤ（Digital Versatile Disc）などの光ディスク、半導体メモリなど、任意の種類の記録媒体を使用することができる。媒体リーダ１０６は、例えば、記録媒体１１３から読み取ったプログラムやデータを、ＲＡＭ１０２やＨＤＤ１０３などの他の記録媒体にコピーする。読み取られたプログラムは、例えば、ＣＰＵ１０１によって実行される。なお、記録媒体１１３は可搬型記録媒体であってもよく、プログラムやデータの配布に用いられることがある。また、記録媒体１１３やＨＤＤ１０３を、コンピュータ読み取り可能な記録媒体と言うことがある。

【００４１】

通信インタフェース１０７は、ネットワーク１１４に接続され、ネットワーク１１４を介して他の情報処理装置と通信する。通信インタフェース１０７は、スイッチやルータなどの有線通信装置に接続される有線通信インタフェースでもよいし、基地局やアクセスポイントなどの無線通信装置に接続される無線通信インタフェースでもよい。

【００４２】

次に、機械学習における訓練データサイズと予測性能について説明する。

第２の実施の形態の機械学習では、既知の事例を示す複数のレコードを含むデータ集合を予め収集しておく。レコードを、サンプル、インスタンス、行、データブロック、単位データなどと言うこともできる。機械学習装置１００または他の情報処理装置が、センサデバイスなどの各種デバイスからネットワーク１１４経由でデータ集合を収集してもよい。収集されるデータ集合は、「ビッグデータ」と言われるサイズの大きなものであってもよい。各レコードは、通常、１以上の説明変数の値と１つの目的変数の値とを含む。例えば、商品の需要予測を行う機械学習では、気温や湿度など商品需要に影響を与える要因を説明変数とし、商品需要量を目的変数とした実績データを収集する。

【００４３】

機械学習装置１００は、収集されたデータ集合の中から一部のレコードを訓練データとしてサンプリングし、訓練データを用いてモデルを学習する。モデルは、説明変数と目的変数との間の関係を示し、通常、１以上の説明変数と１以上の係数と１つの目的変数とを含む。モデルは、例えば、線形式、二次以上の多項式、指数関数、対数関数などの各種数式によって表されてもよい。数式の形は、機械学習の前にユーザによって指定されてもよい。係数は、機械学習によって訓練データに基づいて決定される。

【００４４】

学習されたモデルを用いることで、未知の事例の説明変数の値（要因）から、未知の事例の目的変数の値（結果）を予測することができる。例えば、来期の気象予報から来期の商品需要量を予測できる。モデルによって予測される結果は、０以上１以下の確率などの

連続量であってもよいし、ＹＥＳ／ＮＯの２値などの離散値であってもよい。

【００４５】

学習されたモデルに対しては「予測性能」を算出することができる。予測性能は、未知の事例の結果を正確に予測する能力であり、「精度」と言うこともできる。機械学習装置１００は、収集されたデータ集合の中から訓練データ以外のレコードをテストデータとしてサンプリングし、テストデータを用いて予測性能を算出する。機械学習装置１００は、テストデータに含まれる説明変数の値をモデルに入力し、モデルが出力する目的変数の値（予測値）とテストデータに含まれる目的変数の値（実績値）とを比較する。なお、学習したモデルの予測性能を検証することを「バリデーション」と言うことがある。

【００４６】

予測性能の指標としては、正答率（Accuracy）、適合率（Precision）、平均二乗誤差（ＭＳＥ）、二乗平均平方根誤差（ＲＭＳＥ）などが挙げられる。例えば、結果がＹＥＳ／ＮＯの２値で表されるとする。また、ｎ件のテストデータのレコードのうち、予測値＝ＹＥＳかつ実績値＝ＹＥＳの件数をＴｐ、予測値＝ＹＥＳかつ実績値＝ＮＯの件数をＦｐ、予測値＝ＮＯかつ実績値＝ＹＥＳの件数をＦｎ、予測値＝ＮＯかつ実績値＝ＮＯの件数をＴｎとする。正答率は予測が当たった割合であり、（Ｔｐ＋Ｔｎ）／ｎと算出される。適合率は「ＹＥＳ」の予測を間違えない確率であり、Ｔｐ／（Ｔｐ＋Ｆｐ）と算出される。平均二乗誤差ＭＳＥは、各事例の実績値をＹと表し予測値をｙと表すと、ｓｕｍ（Ｙ－ｙ）^{2}／ｎと算出される。二乗平均平方根誤差ＲＭＳＥは、（ｓｕｍ（Ｙ－ｙ）^{2}／ｎ）^{1}／^{2}と算出される。ＭＳＥ＝ＲＭＳＥ^{2}である。

【００４７】

ここで、訓練データからモデルを学習する手順（機械学習アルゴリズム）には様々なものが存在する。機械学習装置１００が使用する機械学習アルゴリズムは、ユーザから指定されてもよいし、機械学習装置１００が所定の評価方法に従って選択するようにしてもよい。機械学習装置１００が使用できる機械学習アルゴリズムの数は、数十～数百程度あってもよい。機械学習アルゴリズムの一例として、ロジスティック回帰分析、サポートベクタマシン、ランダムフォレストなどを挙げることができる。

【００４８】

ロジスティック回帰分析は、目的変数ｙの値と説明変数x_1，x_2，…，x_dの値をＳ字曲線にフィッティングする回帰分析である。目的変数ｙおよび説明変数x_1，x_2，…，x_dは、ｌｏｇ（ｙ／（１－ｙ））＝$a_1x_1+a_2x_2+\cdots+a_dx_d+b$の関係を満たすと仮定される。$a_1$，$a_2$，…，$a_d$，ｂは係数であり、回帰分析によって決定される。

【００４９】

サポートベクタマシンは、空間に配置されたレコードの集合を、２つのクラスに最も明確に分割するような境界面を算出する機械学習アルゴリズムである。境界面は、各クラスとの距離（マージン）が最大になるように算出される。

【００５０】

ランダムフォレストは、複数の単位データを適切に分類するためのモデルを生成する機械学習アルゴリズムである。ランダムフォレストでは、データ集合からレコードをランダムにサンプリングする。説明変数の一部をランダムに選択し、選択した説明変数の値に応じてサンプリングしたレコードを分類する。説明変数の選択とレコードの分類を繰り返すことで、複数の説明変数の値に基づく階層的な決定木を生成する。レコードのサンプリングと決定木の生成を繰り返すことで複数の決定木を取得し、それら複数の決定木を合成することで、レコードを分類するための最終的なモデルを生成する。

【００５１】

あるデータ集合に１つの機械学習アルゴリズムを適用する場合、訓練データとしてサンプリングするレコードの数（訓練データサイズ）が大きいほど予測性能は高くなる。

図３は、訓練データサイズと予測性能の関係例を示すグラフである。

【００５２】

曲線３１は、モデルの予測性能と訓練データサイズとの間の関係を示す。訓練データサ

イズｓ$_1$，ｓ$_2$，ｓ$_3$，ｓ$_4$，ｓ$_5$の間の大小関係は、ｓ$_1$＜ｓ$_2$＜ｓ$_3$＜ｓ$_4$＜ｓ$_5$である。例えば、ｓ$_2$はｓ$_1$の２倍または４倍であり、ｓ$_3$はｓ$_2$の２倍または４倍であり、ｓ$_4$はｓ$_3$の２倍または４倍であり、ｓ$_5$はｓ$_4$の２倍または４倍である。

【００５３】

曲線３１が示すように、訓練データサイズがｓ$_2$の場合の予測性能はｓ$_1$の場合よりも高い傾向にある。同様に、訓練データサイズがｓ$_3$の場合の予測性能はｓ$_2$の場合よりも高い傾向にある。訓練データサイズがｓ$_4$の場合の予測性能はｓ$_3$の場合よりも高い傾向にある。訓練データサイズがｓ$_5$の場合の予測性能はｓ$_4$の場合よりも高い傾向にある。このように、訓練データサイズが大きくなるほど予測性能も高くなる傾向にある。ただし、予測性能が低いうちは、訓練データサイズの増加に応じて予測性能が大きく上昇する。一方で、予測性能には上限があり、予測性能が上限に近づくと、訓練データサイズの増加量に対する予測性能の上昇量の比は逓減する。すなわち、曲線３１は、訓練データサイズの増加に応じて、ある上限に漸近するように予測性能が増加することを示している。

【００５４】

このような訓練データサイズと予測性能との間の関係は、使用する機械学習アルゴリズムによって異なり、収集したデータ集合の性質（データ集合の種類）によっても異なる。このため、曲線３１に示すような予測性能の上限や各訓練データサイズにおける予測性能を、機械学習を開始する前に推定することは容易でない。

【００５５】

次に、予測性能の測定値の信頼性について説明する。

図４は、予測性能の測定値の分散例を示すグラフである。

ある訓練データサイズのもとで学習されたモデルの予測性能の測定値は、機械学習アルゴリズムとデータ集合の性質とから決まる期待値から乖離するリスクがある。すなわち、同じデータ集合を使用しても、訓練データおよびテストデータとして選択するレコードの偶然性によって、予測性能の測定値にばらつきが生じる。測定値の「ばらつき」は、分散や標準偏差などと解釈することもできる。分散は、訓練データサイズが小さいほど大きく、訓練データサイズが大きいほど小さくなる傾向にある。また、分散は、テストデータサイズが小さいほど大きく、テストデータサイズが大きいほど小さくなる傾向にある。

【００５６】

グラフ３２は、訓練データサイズと予測性能との間の関係を示す。ここでは、同じ機械学習アルゴリズムおよび同じデータ集合を用いて、訓練データサイズ１つ当たり５０回ずつモデルの生成および予測性能の測定を行っている。また、テストデータサイズは、訓練データサイズの２分の１または４分の１とするなど、訓練データサイズに比例するようにして訓練データサイズに従属させている。グラフ３２は、１つの訓練データサイズにつき５０個の予測性能の測定値をプロットしたものである。なお、グラフ３２では、予測性能の指標として、値が大きいほど予測性能が高いことを示す正答率を用いている。

【００５７】

グラフ３２では、訓練データサイズ＝１００の場合の予測性能の測定値は、約０．５８～０．６８であり広範囲に広がっている。訓練データサイズ＝５００の場合の予測性能の測定値は、約０．６９～０．７５であり、訓練データサイズ＝１００の場合よりも範囲が狭くなっている。以降、訓練データサイズが大きくなるに従って測定値の範囲は狭くなる。訓練データサイズが十分に大きくなると、測定値は約０．７６に収束している。

【００５８】

以下では、予測性能の測定値の分散について更に検討する。

まず、バイアス・バリアンス分解について説明する。バイアス・バリアンス分解は、ある機械学習アルゴリズムの良否を評価するために用いられることがある。バイアス・バリアンス分解では、ロス（損失）とバイアスとバリアンスという３つの指標が用いられる。ロス＝バイアスの二乗＋バリアンスという関係が成立する。

【００５９】

ロスは、機械学習によって生成されるモデルが予測を外す度合いを示す指標である。ロ

スの種類には０－１ロスや二乗ロスなどがある。０－１ロスは、予測に成功すれば０を付与し予測に失敗すれば１を付与することで算出されるロスであり、その期待値は予測が失敗する確率を示す。予測が外れることが少ないほど０－１ロスの期待値は小さく、予測が外れることが多いほど０－１ロスの期待値は大きい。二乗ロスは、予測値と真の値との差（予測誤差）の二乗である。予測誤差が小さいほど二乗ロスは小さく、予測誤差が大きいほど二乗ロスは大きい。期待ロス（ロスの期待値）と予測性能とは相互に変換できる。

【００６０】

予測性能が正答率（Accuracy）でありロスが０－１ロスである場合、期待ロス＝１－予測性能である。予測性能が平均二乗誤差（ＭＳＥ）でありロスが二乗ロスである場合、期待ロス＝ＭＳＥである。予測性能が二乗平均平方根誤差（ＲＭＳＥ）でありロスが二乗ロスである場合、期待ロス＝ＲＭＳＥの二乗である。

【００６１】

バイアスは、機械学習によって生成されるモデルの予測値が真の値に対して偏る程度を示す指標である。バイアスが小さいほど精度の高いモデルであると言うことができる。バリアンスは、機械学習によって生成されるモデルの予測値がばらつく程度を示す指標である。バリアンスが小さいほど精度の高いモデルであると言うことができる。ただし、バイアスとバリアンスの間にはトレードオフの関係があることが多い。

【００６２】

次数の小さい多項式など複雑性の低いモデル（表現力の低いモデルと言うこともできる）では、モデルの係数をどの様に調整しても、複数のレコードの全てについて真の値に近い予測値を出力するようにすることは難しい。すなわち、複雑性の低いモデルを用いると複雑な事象を表現できない。よって、複雑性の低いモデルのバイアスは大きくなる傾向にある。この点、次数の大きい多項式など複雑性の高いモデル（表現力の高いモデルと言うこともできる）では、モデルの係数を適切に調整することで、複数のレコードの全てについて真の値に近い予測値を出力することができる余地がある。よって、複雑性の高いモデルのバイアスは小さくなる傾向にある。

【００６３】

一方で、複雑性の高いモデルでは、訓練データとして使用するレコードの特徴に過度に依存したモデルが生成されるという過学習が生じるリスクがある。過学習によって生成されたモデルは、他のレコードについて適切な予測値を出力できないことが多い。例えば、ｄ次の多項式を用いると、ｄ＋１個のレコードについて真の値と完全に一致する予測値を出力するモデル（残差が０のモデル）を生成できる。

【００６４】

しかし、あるレコードについて残差が０になるモデルは、通常は過度に複雑なモデルであり、他のレコードについて予測誤差が著しく大きい予測値を出力してしまうリスクが高くなる。よって、複雑性の高いモデルのバリアンスは大きくなる傾向にある。この点、複雑性の低いモデルでは、予測誤差が著しく大きい予測値を出力してしまうリスクは低く、バリアンスは小さくなる傾向にある。このように、ロスの成分としてのバイアスとバリアンスは、モデルを生成する機械学習アルゴリズムの特性に依存している。

【００６５】

次に、ロスとバイアスとバリアンスの形式的定義を説明する。ここでは、二乗ロスをバイアスとバリアンスに分解する例について説明する。

同一のデータ集合からｍ個の訓練データD_k（ｋ＝１，２，…，ｍ）が抽出され、ｍ個のモデルが生成されたとする。また、上記のデータ集合からｉ個のレコードを含むテストデータＴが抽出されたとする。ｉ番目のレコード（テストケースと言うこともできる）は、説明変数の値X_iと目的変数の真の値Y_iとを含む（ｉ＝１，２，…，ｎ）。ｋ番目のモデルからは説明変数の値X_iに対して目的変数の予測値y_{ik}が算出される。

【００６６】

すると、ｋ番目のモデルとｉ番目のレコードとの間で算出される予測誤差e_{ik}は$e_{ik}=Y_i-y_{ik}$と定義され、そのロス（ここでは二乗ロス）はe_{ik}^2と定義される。ｉ番

目のレコードに対しては、バイアスB_iとバリアンスV_iとロスL_iが定義される。バイアスB_iは$B_i = E_D[e_{ik}]$と定義される。$E_D[]$はm個の訓練データの間の平均値（期待値）を表す。バリアンスV_iは$V_i = V_D[e_{ik}]$と定義される。$V_D[]$はm個の訓練データの間の分散を表す。ロスL_iは$L_i = E_D[e_{ik}^2]$と定義される。前述のロスとバイアスとバリアンスの間の関係から$L_i = B_i^2 + V_i$が成立する。

【００６７】

テストデータＴ全体に対しては、期待バイアスＥＢ２と期待バリアンスＥＶと期待ロスＥＬが定義される。期待バイアスＥＢ２は$EB2 = E_x[B_i^2]$と定義される。$E_x[]$はｎ個のレコードの間の平均値（期待値）を表す。期待バリアンスＥＶは$EV = E_x[V_i]$と定義される。期待ロスＥＬは$EL = E_x[L_i]$と定義される。前述のロスとバイアスとバリアンスの間の関係からＥＬ＝ＥＢ２＋ＥＶが成立する。

【００６８】

バイアス・バリアンス分解の考え方を応用して、予測性能の測定値に生じる分散を推定することができる。測定値の分散は、次の数式によって近似される。ＶＬ＝Ｃ×（ＥＬ＋ＥＢ２）×（ＥＬ－ＥＢ２）。ＶＬは訓練データサイズｓにおける予測性能の測定値の分散を表す。Ｃは定数である。ＥＬは訓練データサイズｓにおける期待ロスを表す。ＥＢ２は期待バイアスを表す。以下、この数式の意味について説明を加える。

【００６９】

図５は、予測性能の期待ロスおよび期待バイアスの例を示すグラフである。

曲線３３は、訓練データサイズとロスの推定値との間の関係を示すロス曲線である。図３では縦軸が予測性能であるのに対し、図５では縦軸がロスに変換されている。前述のように予測性能とロスは、予測性能の指標とロスの指標に応じて相互に変換可能である。曲線３３は、訓練データサイズの増加に応じてロスが単調に減少し一定の下限ロスに漸近する非線形曲線である。訓練データサイズが小さいうちはロスの減少量が大きく、訓練データサイズが大きくなるとロスの減少量が小さくなる。

【００７０】

訓練データサイズs_pにおける曲線３３上の点のロス（ロス＝０から曲線３３上の点までの距離）は、訓練データサイズs_pの期待ロスEL_pに相当する。曲線３３によって特定される下限ロスは、図３の曲線３１によって特定される予測性能の上限に対応しており、０より大きい値である。例えば、予測性能の上限をｃとおくと、予測性能が正答率である場合、下限ロスは１－ｃとなる。予測性能が平均二乗誤差（ＭＳＥ）である場合、下限ロスはｃとなる。予測性能が二乗平均平方根誤差（ＲＭＳＥ）である場合、下限ロスはc^2となる。下限ロスは、この機械学習アルゴリズムにとっての期待バイアスＥＢ２に相当する。訓練データサイズが十分大きくなると、機械学習に用いる訓練データの特徴がデータ集合の特徴に一致し、期待バリアンスが０に近づくためである。

【００７１】

期待ロスEL_pと期待バイアスＥＢ２の差は、訓練データサイズs_pにおけるギャップと言うことができる。ギャップは、訓練データサイズを大きくすることでその機械学習アルゴリズムがロスを低減できる余地を表している。ギャップは、図３の曲線３１上の点と予測性能の上限との間の距離に対応し、訓練データサイズを大きくすることでその機械学習アルゴリズムが予測性能を改善できる余地を表しているとも言える。ギャップは、訓練データサイズs_pにおける期待バリアンスの影響を受ける。

【００７２】

次に、予測性能の測定値の分散を示す数式の数学的根拠について説明する。

（ａ）問題の形式的な記述

同一のデータ集合からｍセットの訓練データD_1，D_2，…，D_mとテストデータＴが抽出されたとする。ある機械学習アルゴリズムに訓練データD_kを与えて学習されたモデルをf_kとする（ｋ＝１，２，…，ｍ）。テストデータＴをレコード$<Y_i, X_i>$の集合とする（ｉ＝１，２，…，ｎ）。X_iは説明変数の値（入力値）であり、Y_iは入力値X_iに対応する目的変数の既知の値（真値）である。入力値X_iに対してモデルf_kが予測した

値（予測値）を$y_{ik}=f_k(X_i)$とする。入力値X_iに対するモデルf_kによる予測の誤差は$e_{ik}=Y_i-y_{ik}$と定義される。テストデータＴに含まれるレコードの数、すなわち、テストデータＴのサイズはｎである。以下では主に、ｉ，ｊはテストデータＴのレコードを識別する添え字、ｋはモデルを識別する添え字として使用する。

【００７３】

機械学習アルゴリズムが回帰を目的とする場合、予測値は連続量であり、ロスの指標として数式（１）の二乗ロスが用いられることが多い。この二乗ロスをテストデータＴの全てのレコードについて平均したものが数式（２）のＭＳＥ（平均二乗誤差）である。

【００７４】

【数１】

$$\mathrm{loss}_{\mathrm{sq}}(e)=e^2 \tag{1}$$

【００７５】

【数２】

$$\mathrm{MSE}=\mathrm{E}_{\mathrm{X}}[\mathrm{loss}_{\mathrm{sq}}(e_{ik})]=\mathrm{E}_{\mathrm{X}}[e_{ik}^2] \tag{2}$$

【００７６】

ここで、Ｅ［・］は期待値を求める演算子であり、Ｖ［・］は分散を求める演算子である。Ｅ［・］，Ｖ［・］に付加する添え字Ｘは、この演算子がテストデータＴの複数のレコードの間の演算であることを示す。Ｅ［・］，Ｖ［・］に付加する添え字Ｍは、この演算子が複数のモデルの間の演算であることを示す。すなわち、E_{X}［・］はテストデータＴの複数のレコードの間で平均化した期待値を示し、E_{M}［・］は複数のモデルの間で平均化した期待値を示す。V_{X}［・］はテストデータＴの複数のレコードの間の分散を示し、V_{M}［・］は複数のモデルの間の分散を示す。また、ｃｏｖ（・，・）は共分散を求める共分散関数であり、ｃｏｒ（・，・）は相関係数を求める相関係数関数である。ｃｏｖ（・，・），ｃｏｒ（・，・）にも添え字Ｘ，Ｍが付加される。

【００７７】

機械学習アルゴリズムが二値分類を目的とする場合、予測値は｛－１，１｝のような二値の離散値であり、ロスの指標として数式（３）の０－１ロスが用いられることが多い。この０－１ロスをテストデータＴの全てのレコードについて平均して１から引いたものが、数式（４）の正答率（Accuracy）である。

【００７８】

【数３】

$$\mathrm{loss}_{01}(e)=\begin{cases}0 & \text{if } e=0\\ 1 & \text{otherwise}\end{cases} \tag{3}$$

【００７９】

【数４】

$$\mathrm{Accuracy}=1-\mathrm{E}_{\mathrm{X}}[\mathrm{loss}_{01}(e_{ik})]=1-\mathrm{E}_{\mathrm{X}}[e_{ik}^2] \tag{4}$$

【００８０】

ＭＳＥは値が小さいほど予測性能が高いことを示し、正答率は値が大きいほど予測性能が高いことを示す。ただし、両者ともテストデータＴ全体について平均化したロスがモデルの予測性能の良否を表している点で共通しており、モデルロスと言うことができる。モデルf_kに対するモデルロスML_kは数式（５）の通りである。予測性能の分散は、数式

（６）のように複数のモデルの間のモデルロスの分散として表される。

【００８１】

【数５】

$$ML_k = \mathrm{E_X}[e_{ik}^2] \tag{5}$$

【００８２】

【数６】

$$\mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] = \mathrm{V_M}[ML_k] \tag{6}$$

【００８３】

（ｂ）バイアス・バリアンス分解

モデルによる予測で生じるロスはバイアスとバリアンスに分解できる。バイアスはモデルの予測値の偏りを示す量である。バイアスが低いモデルほど正確なモデルであると言える。表現力の低いモデル（調整可能な係数が少ないような複雑性の低いモデルなど）はバイアスが高くなる傾向にある。バリアンスはモデルの予測値のばらつきを示す量である。バリアンスが低いほど正確なモデルであると言える。表現力の高いモデル（調整可能な係数が多いような複雑性の高いモデルなど）はバリアンスが高くなる傾向にある。表現力の高いモデルには、訓練データに過剰適合するという過学習のリスクがある。

【００８４】

テストデータＴの入力値Ｘ$_i$に対するロスＬ$_i$、バイアスＢ$_i$およびバリアンスＶ$_i$は、数式（７）～（９）のように定義される。ロスＬ$_i$は複数のモデルの間の二乗誤差の期待値であり、バイアスＢ$_i$は複数のモデルの間の誤差の期待値であり、バリアンスＶ$_i$は複数のモデルの間の誤差の分散である。ロスＬ$_i$とバイアスＢ$_i$とバリアンスＶ$_i$との間には、数式（１０）の関係（バイアス・バリアンス分解）が成立する。

【００８５】

【数７】

$$L_i = \mathrm{E_M}[e_{ik}^2] \tag{7}$$

【００８６】

【数８】

$$B_i = \mathrm{E_M}[e_{ik}] \tag{8}$$

【００８７】

【数９】

$$V_i = \mathrm{V_M}[e_{ik}] \tag{9}$$

【００８８】

【数１０】

$$L_i = V_i + B_i^2 \tag{10}$$

【００８９】

様々な入力値Ｘ$_i$に対するロスＬ$_i$の期待値を期待ロスＥＬ、バイアスＢ$_i$の二乗の期待値を期待バイアスＥＢ２、バリアンスＶ$_i$の期待値を期待バリアンスＥＶとする。期待ロスＥＬ、期待バイアスＥＢ２、期待バリアンスＥＶは、数式（１１）～（１３）のように

定義される。期待ロスＥＬと期待バイアスＥＢ２と期待バリアンスＥＶとの間には、数式（１４）の関係（バイアス・バリアンス分解）が成立する。

【００９０】

【数１１】

$$EL = \mathrm{E_X}[L_i] \tag{11}$$

【００９１】

【数１２】

$$EB2 = \mathrm{E_X}[B_i^2] \tag{12}$$

【００９２】

【数１３】

$$EV = \mathrm{E_X}[V_i] \tag{13}$$

【００９３】

【数１４】

$$EL = EV + EB2 \tag{14}$$

【００９４】

ここでの目的は、ＥＬ，ＥＢ２，ＥＶとモデルロスの分散との間の関係を導出することである。なお、期待ロスＥＬとモデルロスＭＬ$_k$の期待値とは、数式（１５）に示すように等価である。一方、ロスＬ$_i$の分散とモデルロスＭＬ$_k$の分散とは等価でない。以下では、予測性能の分散を推定する数式を次の流れで導出する。第１に、ロスの分散をバイアスとバリアンスで記述する。第２に、モデルロスの分散をインスタンス成分と相互作用成分に分解する。第３に、インスタンス成分を算出する。第４に、相互作用成分を算出する。第５に、モデルロスの分散をバイアスとバリアンスで記述する。

【００９５】

【数１５】

$$EL = \mathrm{E_X}[L_i] = \mathrm{E_X}[\mathrm{E_M}[e_{ik}^2]] = \mathrm{E_M}[\mathrm{E_X}[e_{ik}^2]] = \mathrm{E_M}[ML_k] \tag{15}$$

【００９６】

（ｃ）ロスの分散をバイアスとバリアンスで記述

テストデータＴの入力値Ｘ$_i$を固定して複数のモデルの誤差を並べた誤差ベクトルを考える。誤差ｅを確率変数とみなしてその分布が正規分布に従うと仮定すると、複数のモデルの間のロスの分散は数式（１６）のように定義され、バイアスＢ$_i$とバリアンスＶ$_i$の組またはロスＬ$_i$とバイアスＢ$_i$の組によって記述することができる。数式（１６）の１行目から２行目への変形では、数式（１７）に示す統計学上の性質（確率変数の４乗の期待値）が利用されている。数式（１７）においてＸは確率変数であり、Ｓは歪度であり、Ｋは尖度である。正規分布の場合はＳ＝０かつＫ＝３である。

【００９７】

【数１６】

$$\begin{aligned} \mathrm{V_M}[e_{ik}^2] &= \mathrm{E_M}[e_{ik}^4] - (\mathrm{E_M}[e_{ik}^2])^2 \\ &= 2V_i^2 + 4V_iB_i^2 = 2L_i^2 - 2B_i^4 \end{aligned} \tag{16}$$

【００９８】
【数１７】

$$\mathrm{E}[X^4] = K(\mathrm{V}[X])^2 + 4S(\mathrm{V}[X])^{1.5}\mathrm{E}[X] + 6\mathrm{V}[X](\mathrm{E}[X])^2 + (\mathrm{E}[X])^4 \quad (17)$$

【００９９】
（ｄ）モデルロスの分散をインスタンス成分と相互作用成分に分解
分散の基本的性質から、予測性能の分散（複数のモデルの間のモデルロスの分散）について数式（１８）が成立する。これをｎ×ｎ行列の成分の平均と考えると、ｉ＝ｊである対角成分は入力値Ｘ ｉに対するロスの分散を表しており、その相関係数は１になる。一方、ｉ≠ｊである非対角成分の相関係数は異なる入力値の間のロスの相関を表している。異なる入力値に対する誤差の発生状況は共通点が少ないため、その相関係数の絶対値は十分に小さくなることが多く、予測性能の高いモデルほどその相関係数は０に近づく。対角成分と非対角成分とは性質が異なるため、数式（１９）のように両者を分離して考える。
【０１００】
【数１８】

$$\begin{aligned}\mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] &= \mathrm{E_{X}}_i[\mathrm{E_{X}}_j[\mathrm{cov_M}(e_{ik}^2, e_{jk}^2)]] \\ &= \mathrm{E_{X}}_i[\mathrm{E_{X}}_j[\mathrm{cor_M}(e_{ik}^2, e_{jk}^2)\cdot(\mathrm{V_M}[e_{ik}^2])^{0.5}(\mathrm{V_M}[e_{jk}^2])^{0.5}]]\end{aligned} \quad (18)$$

【０１０１】
【数１９】

$$\begin{aligned}&\mathrm{E_{X}}_i[\mathrm{E_{X}}_j[\mathrm{cov_M}(e_{ik}^2, e_{jk}^2)]] \\ &= \frac{n}{n^2}\mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j=i}[\mathrm{cov_M}(e_{ik}^2, e_{jk}^2)]] + \frac{n(n-1)}{n^2}\mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[\mathrm{cov_M}(e_{ik}^2, e_{jk}^2)]] \\ &= \frac{1}{n}\mathrm{E_X}[\mathrm{V_M}[e_{ik}^2]] + \frac{n-1}{n}\mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[\mathrm{cov_M}(e_{ik}^2, e_{jk}^2)]]\end{aligned} \quad (19)$$

【０１０２】
数式（１９）では、モデルロスの分散を第１項のインスタンス成分と第２項の相互作用成分とに分解している。第１項はロスの分散の期待値を表しており、モデルロスの分散の大部分を占めることが多い。第２項は異なる入力値の間の共分散の期待値を表しており、モデルロスの分散に対する寄与は小さいことが多い。第１項はテストデータＴのサイズｎに反比例するため、テストデータＴのレコードを増やすことでモデルロスの分散を低減できる。ただし、第２項が存在することから低減効果には限界がある。
（ｅ）インスタンス成分を算出
数式（１９）の第１項について検討する。上記の数式（１６）より数式（２０）が成立する。ここで、数式（２０）の第１項と第２項を算出するために幾つかの仮定をおく。多くの機械学習アルゴリズムは不偏推定量を出力するようにモデルを学習することから、数式（２１）のように誤差の期待値が０になるという仮定をおく。数式（２１）からバイアスＢ ｉについて数式（２２）の性質が導出される。
【０１０３】
【数２０】

$$\mathrm{E_X}[\mathrm{V_M}[e_{ik}^2]] = 2\mathrm{E_X}[L_i^2] - 2\mathrm{E_X}[B_i^4] \quad (20)$$

【０１０４】
【数２１】

$$\mathrm{E_X}[e_{ik}] = 0 \tag{21}$$

【０１０５】
【数２２】

$$\mathrm{E_X}[B_i] = 0 \tag{22}$$

【０１０６】
確率分布の中には、訓練データサイズや訓練データのサンプリング方法に依存して期待値や分散が変化することはあっても、確率分布の形状を示す歪度や尖度は変化しない（または、変化が非常に緩やかである）ものがあると仮定する。具体的には、入力値Ｘ$_i$に対する複数のモデルの間の誤差の分布は正規分布を形成し、尖度＝３かつ歪度＝０になることを仮定する。また、バイアスＢ$_i$の分布の尖度Ｋ１は変化しないことを仮定する。バイアスＢ$_i$の分布の尖度Ｋ１は、数式（２３）のように定義される。数式（２３）と上記の数式（１２）から数式（２４）が算出される。
【０１０７】
【数２３】

$$K1 = \frac{\mathrm{E_X}[B_i^4]}{(\mathrm{E_X}[B_i^2])^2} \tag{23}$$

【０１０８】
【数２４】

$$\mathrm{E_X}[B_i^4] = K1 \cdot EB2^2 \tag{24}$$

【０１０９】
また、モデルｆ$_k$に対する複数の入力値の間の誤差の分布の尖度Ｋ２は、モデル間で共通でありかつ変化しないことを仮定する。尖度Ｋ２は数式（２５）のように定義される。Ｋ１，Ｋ２の値はそれぞれ３～１０の範囲内であることが多く、両者は近いことが多い。
【０１１０】
【数２５】

$$K2 = \frac{\mathrm{E_M}[\mathrm{E_X}[e_{ik}^4]]}{\mathrm{E_M}[(\mathrm{E_X}[e_{ik}^2])^2]} \tag{25}$$

【０１１１】
数式（２５）から数式（２６）が導出される。数式（２６）を数式（１８），（１９）に代入することで数式（２７）が算出される。ここで、尖度Ｋ２はサイズｎより十分に小さいため、１－Ｋ２／ｎは１に近似される。数式（２０），（２３）を数式（１８），（１９）に代入することで数式（２８）が算出される。数式（２８）から数式（２７）を減算して数式（２９）が算出される。そして、数式（２０），（２４），（２９）から数式（３０）が算出される。これが、数式（１９）の第１項の主要成分である。
【０１１２】
【数２６】

$$\begin{aligned}\mathrm{E_X}[\mathrm{V_M}[e_{ik}^2]] &= \mathrm{E_X}[\mathrm{E_M}[e_{ik}^4]] - \mathrm{E_X}[(\mathrm{E_M}[e_{ik}^2])^2] \\ &= K2 \cdot \mathrm{E_M}[(\mathrm{E_X}[e_{ik}^2])^2] - K2(\mathrm{E_M}[\mathrm{E_X}[e_{ik}^2]])^2 \\ &\quad + K2(\mathrm{E_M}[\mathrm{E_X}[e_{ik}^2]])^2 - \mathrm{E_X}[(\mathrm{E_M}[e_{ik}^2])^2] \qquad (26) \\ &= K2 \cdot \mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] + K2 \cdot EL^2 - \mathrm{E_X}[L_i^2]\end{aligned}$$

【０１１３】
【数２７】

$$\left(1 - \frac{K2}{n}\right)\mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] \approx \mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]]$$
$$= \frac{1}{n}(K2 \cdot EL^2 - \mathrm{E_X}[L_i^2]) + \frac{n-1}{n}\mathrm{E_X}_i[\mathrm{E_X}_{j,j\neq i}[\mathrm{cov_M}(e_{ik}^2, e_{jk}^2)]] \quad (27)$$

【０１１４】
【数２８】

$$\begin{aligned}&\mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] \qquad (28) \\ &= \frac{1}{n}(2\mathrm{E_X}[L_i^2] - 2K1 \cdot EB2^2) + \frac{n-1}{n}\mathrm{E_X}_i[\mathrm{E_X}_{j,j\neq i}[\mathrm{cov_M}(e_{ik}^2, e_{jk}^2)]]\end{aligned}$$

【０１１５】
【数２９】

$$\mathrm{E_X}[L_i^2] = \frac{1}{3}K2 \cdot EL^2 + \frac{2}{3}K1 \cdot EB2^2 \qquad (29)$$

【０１１６】
【数３０】

$$\mathrm{E_X}[\mathrm{V_M}[e_{ik}^2]] = \frac{2}{3}K2 \cdot EL^2 - \frac{2}{3}K1 \cdot EB2^2 \qquad (30)$$

【０１１７】
（ｆ）相互作用成分を算出
不動点Ｃｏｒ１ｖを数式（３１）のように定義する。不動点Ｃｏｒ１ｖは、訓練データサイズを変化させても値が変化しないかまたは非常に緩やかに変化することが多いため、ここでは訓練データサイズに依存しないと仮定する。不動点Ｃｏｒ１ｖの値は０．００１～０．１程度であることが多い。
【０１１８】
【数３１】

$$\begin{aligned}Cor1v &= \mathrm{V}_{i,j,j\neq i}[\mathrm{cor_M}(e_{ik}, e_{jk})] \qquad (31) \\ &= \mathrm{E_X}_i[\mathrm{E_X}_{j,j\neq i}[\mathrm{cor_M}(e_{ik}, e_{jk})^2]] - (\mathrm{E_X}_i[\mathrm{E_X}_{j,j\neq i}[\mathrm{cor_M}(e_{ik}, e_{jk})]])^2\end{aligned}$$

【０１１９】
ここで、数式（３２）に示す統計学上の性質（誤差の相関係数の期待値）を利用する。

誤差の期待値が０であるとき、２つの誤差の相関係数の期待値は０に近似する。この性質から数式（３３）が成立し、上記の数式（３１）から数式（３４）が算出される。

【０１２０】

【数３２】

$$\mathrm{E_X}[e_{ik}] = 0 \quad \Rightarrow \quad \mathrm{E_{X}}_i[\mathrm{E_{X}}_j[\mathrm{cor_M}(e_{ik}, e_{jk})]] \approx 0 \qquad (32)$$

【０１２１】

【数３３】

$$\mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[\mathrm{cor_M}(e_{ik}, e_{jk})]] = \frac{n^2 \mathrm{E_{X}}_i[\mathrm{E_{X}}_j[\mathrm{cor_M}(e_{ik}, e_{jk})]] - n}{n(n-1)}$$
$$\approx \frac{-1}{n-1} \qquad (33)$$

【０１２２】

【数３４】

$$\mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[\mathrm{cor_M}(e_{ik}, e_{jk})^2]] \approx Cor1v + \frac{1}{(n-1)^2} \qquad (34)$$

【０１２３】

また、数式（３５）が成立する。数式（３５）の２行目から３行目への変形では、相関係数ｃｏｒ$_{M}$とバリアンスＶ$_{i}$，Ｖ$_{j}$とは互いに独立であることを仮定している。数式（３５）の３行目から４行目への変形では、上記の数式（３４）を利用しており、Ｖ$_{i}$・Ｖ$_{j}$の期待値がＥＶ^{2}に近似することを利用している。数式（３５）の４行目の近似では、テストデータサイズｎが１より十分に大きいため１／（ｎ－１）^{2}を無視している。

【０１２４】

【数３５】

$$\begin{aligned}
&\mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[\mathrm{cov_M}(e_{ik}, e_{jk})^2]] \\
&= \mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[\mathrm{cor_M}(e_{ik}, e_{jk})^2 V_i \cdot V_j]] \\
&= \mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[\mathrm{cor_M}(e_{ik}, e_{jk})^2]] \cdot \mathrm{E_{X}}_i[\mathrm{E_{X}}_{j,j\neq i}[V_i \cdot V_j]] \qquad (35) \\
&\approx \left(Cor1v + \frac{1}{(n-1)^2}\right) EV^2 \approx Cor1v \cdot EV^2 \approx Cor1v(EL - EB2)^2
\end{aligned}$$

【０１２５】

ここで、数式（３６）に示す統計学上の性質（共分散の二乗と二乗の共分散の関係）を利用する。確率変数Ｘ，Ｙの結合確率が二次元正規分布に従うならば数式（３６）が成立する。誤差の分散が正規分布に従うため、数式（３６）を利用して数式（３７）が算出される。また、数式（３８）が成立する。数式（３８）の１行目から２行目への変形では、共分散ｃｏｖ$_{M}$とバイアスＢ$_{i}$，Ｂ$_{j}$は概ね独立であることを仮定している。数式（３８）の２行目の近似では、Ｂ$_{i}$Ｂ$_{j}$の期待値はバイアスＢ$_{i}$の期待値の二乗に近似しその結果０に近似するという性質を利用している。数式（３５），（３８）を数式（３７）に代入することで数式（３９）が算出される。これが、数式（１９）の第２項の主要成分である。

【０１２６】

【数３６】

$$\mathrm{cov}(X^2,Y^2)=2\mathrm{cov}(X,Y)^2+4\mathrm{E}[X]\mathrm{E}[Y]\mathrm{cov}(X,Y) \qquad (36)$$

【０１２７】
【数３７】

$$\begin{aligned}&\mathrm{E}_{\mathrm{X}i}[\mathrm{E}_{\mathrm{X}j,j\neq i}[\mathrm{cov}_{\mathrm{M}}(e_{ik}^2,e_{jk}^2)]] \qquad (37)\\&=\mathrm{E}_{\mathrm{X}i}[\mathrm{E}_{\mathrm{X}j,j\neq i}[2\mathrm{cov}_{\mathrm{M}}(e_{ik},e_{jk})^2]]+\mathrm{E}_{\mathrm{X}i}[\mathrm{E}_{\mathrm{X}j,j\neq i}[4B_iB_j\mathrm{cov}_{\mathrm{M}}(e_{ik},e_{jk})]]\end{aligned}$$

【０１２８】
【数３８】

$$\begin{aligned}&\mathrm{E}_{\mathrm{X}i}[\mathrm{E}_{\mathrm{X}j,j\neq i}[B_iB_j\mathrm{cov}_{\mathrm{M}}(e_{ik},e_{jk})]]\\&=\mathrm{E}_{\mathrm{X}i}[\mathrm{E}_{\mathrm{X}j,j\neq i}[\mathrm{cov}_{\mathrm{M}}(e_{ik},e_{jk})]]\cdot\mathrm{E}_{\mathrm{X}i}[\mathrm{E}_{\mathrm{X}j,j\neq i}[B_iB_j]]\approx 0 \qquad (38)\end{aligned}$$

【０１２９】
【数３９】

$$\mathrm{E}_{\mathrm{X}i}[\mathrm{E}_{\mathrm{X}j,j\neq i}[\mathrm{cov}_{\mathrm{M}}(e_{ik}^2,e_{jk}^2)]]\approx 2\ Cor1v(EL-EB2)^2 \qquad (39)$$

【０１３０】
（ｇ）モデルロスの分散をバイアスとバリアンスで記述
上記の数式（１８），（１９），（３０），（３９）より数式（４０）の近似式が成立する。尖度Ｋ２は尖度Ｋ１に近似するため、数式（４０）は数式（４１）のように近似される。典型的にはＫ１（ＥＬ＋ＥＢ２）はｃｏｒ１ｖ（ＥＬ－ＥＢ２）より十分に大きいため、数式（４１）は更に数式（４２）のように近似される。尖度Ｋ１は事前には不明であるが、分散の比が判明すれば実用上十分であることも多い。そこで、数式（４２）は比例定数Ｃを用いて数式（４３）のように単純化できる。これにより、予測性能の測定値の分散が、期待ロスＥＬと期待バイアスＥＢ２の差に比例し、かつ、期待ロスＥＬと期待バイアスＥＢ２の和に比例するという数式が導出される。
【０１３１】
【数４０】

$$\begin{aligned}\mathrm{V}_{\mathrm{M}}[\mathrm{E}_{\mathrm{X}}[e_{ik}^2]]\approx\frac{1}{n}\Big(\frac{2}{3}K2\cdot EL^2-\frac{2}{3}K1\cdot EB2^2\Big) \qquad (40)\\+\frac{n-1}{n}\cdot 2\ Cor1v(EL-EB2)^2\end{aligned}$$

【０１３２】
【数４１】

$$\begin{aligned}&\mathrm{V}_{\mathrm{M}}[\mathrm{E}_{\mathrm{X}}[e_{ik}^2]] \qquad (41)\\&\approx\Big(\frac{1}{n}\frac{2}{3}K1(EL+EB2)+\frac{n-1}{n}\cdot 2\ Cor1v(EL-EB2)\Big)(EL-EB2)\end{aligned}$$

【０１３３】
【数４２】

$$\mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] \approx \frac{1}{n}\frac{2}{3}K1(EL+EB2)(EL-EB2) \qquad (42)$$

【０１３４】
【数４３】

$$\begin{aligned}\mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] &\approx C(EL-EB2)(EL+EB2) \qquad (43)\\ &\approx C\cdot EV(EV+2\ EB2)\end{aligned}$$

【０１３５】
予測性能の測定値の分散を示す数式（４１）は、数式（４４）のように変形することができる。テストデータサイズｎに着目すると、数式（４４）の第１項は、予測性能の測定値の分散のうち、テストデータサイズｎの増大に応じて減少するテストデータ依存成分に相当する。一方、数式（４４）の第２項は、予測性能の測定値の分散のうち、テストデータサイズｎの増大によっては減少しない訓練データ依存成分に相当する。このため、数式（４４）は、予測性能の測定値の分散が、テストデータサイズｎの増大によって減少するものの、０より大きい下限が存在することを示している。
【０１３６】
【数４４】

$$\begin{aligned}\mathrm{V_M}[\mathrm{E_X}[e_{ik}^2]] &\approx \frac{1}{n}\frac{2}{3}K1(EL^2-EB2^2)\\ &\quad+\frac{n-1}{n}\cdot 2\ Cor1v(EL-EB2)^2 \qquad (44)\end{aligned}$$

【０１３７】
前述の図４では、テストデータサイズを訓練データサイズの２分の１や４分の１とするなど、テストデータサイズを訓練データサイズに比例するように決定していた。しかし、このようなテストデータサイズの決定方法は、予測性能の測定値の信頼性と予測性能の測定の負荷とを両立させる観点から、改善の余地がある。テストデータサイズが小さ過ぎると、予測性能の測定値が有する潜在的な分散が大きくなり、算出される測定値の信頼性が低下する。一方、テストデータサイズが大き過ぎると、予測性能の測定値の分散があまり減少せず、測定値の信頼性の向上にあまり寄与しないにもかかわらず、テスト処理を無駄に繰り返すことになり、テスト処理の負荷が増大する。
【０１３８】
この点、数式（４４）が示す予測性能の測定値の分散とテストデータサイズｎとの間の対応関係によれば、測定値の信頼性とテスト負荷とを両立させるような効率的なテストデータサイズｎが存在することになる。そこで、第２の実施の形態の機械学習装置１００は、数式（４４）に基づいて、適切なテストデータサイズを決定する。
【０１３９】
データ集合と機械学習アルゴリズムが特定されると、機械学習装置１００は、数式（４４）の尖度Ｋ１と不動点Ｃｏｒ１ｖと期待バイアスＥＢ２を決定する。これにより、機械学習装置１００は、テストデータサイズｎと期待ロスＥＬを引数として有し、予測性能の測定値の分散を推定する分散関数ｆ（ｎ，ＥＬ）を生成する。尖度Ｋ１と不動点Ｃｏｒ１ｖと期待バイアスＥＢ２は、訓練データサイズに依存しないパラメータである。そのため、データ集合と機械学習アルゴリズムが同じであれば、訓練データサイズが異なっても、同じ分散関数を用いて予測性能の測定値の分散を推定することができる。
【０１４０】

あるデータ集合と機械学習アルゴリズムの組に対する尖度Ｋ１と不動点Ｃｏｒ１ｖと期待バイアスＥＢ２を決定するには、前述のように、ｍセットの訓練データと１セットのテストデータの間の網羅的な誤差を示す誤差プロファイルを用意することが好ましい。そこで、機械学習装置１００は、予測性能を測定したいモデルの訓練データサイズよりも十分に小さい訓練データを、同一のデータ集合からｍセット抽出し、ｍセットの訓練データを用いて機械学習によりｍ個のモデルを生成する。また、機械学習装置１００は、十分に小さいテストデータを当該データ集合から抽出し、テストデータに含まれる複数のレコードとｍ個のモデルとの間で網羅的に誤差を算出する。

【０１４１】

例えば、予測性能を測定したいモデルの訓練データサイズが１００万レコードであるとする。この場合、誤差プロファイル生成のための訓練データの個数を１０セットとし、各訓練データのサイズを１万レコードとする。また、テストデータサイズを、訓練データサイズの２分の１である５０００レコードとする。これにより、１０個のモデルとテストデータの５０００レコードとの間で、１０×５０００個の誤差が算出される。機械学習装置１００は、この誤差プロファイルを用いて、数式（２３）の尖度Ｋ１と、数式（３１）の不動点Ｃｏｒ１ｖと、数式（１２）の期待バイアスＥＢ２を算出する。

【０１４２】

分散関数ｆ（ｎ，ＥＬ）が生成されると、機械学習装置１００は、予測性能を測定したいモデルに対応する期待ロスＥＬを分散関数に代入する。期待ロスＥＬは、図５に示すように、データ集合と機械学習アルゴリズムが同じでも訓練データサイズに応じて変化する。そのため、対象のモデルに対応する期待ロスＥＬを使用することになる。

【０１４３】

ある訓練データサイズに対応する期待ロスＥＬは、測定せずに与えられることもあるし対象のモデルから測定して求めることもある。測定しない場合として、データ集合および機械学習アルゴリズムが同一であり訓練データサイズが異なる複数のモデルの予測性能が、既に測定済みである場合が考えられる。その場合、回帰分析などの統計的方法により、それら測定値から未知の期待ロスＥＬを推定することが考えられる。未知の期待ロスＥＬの推定には、図３や図５の非線形曲線を利用することができる。

【０１４４】

対象のモデルから測定する場合、例えば、機械学習装置１００は、誤差プロファイルの生成に使用した小さなテストデータを対象のモデルに入力し、テストデータに含まれる複数のレコードに対応する誤差を算出する。そして、機械学習装置１００は、それら誤差から数式（１１）の期待ロスＥＬを算出する。例えば、５０００レコードのテストデータから５０００個の誤差が算出され、期待値としての期待ロスＥＬが算出される。

【０１４５】

上記の方法で測定される期待ロスＥＬは、対象のモデルが大きい訓練データサイズで学習されているため、当該大きい訓練データサイズに対応した測定値になる。ただし、小さいテストデータを使用するため、大きいテストデータを使用して測定される本来の期待ロスＥＬと比較すると、測定値の分散が大きくなる。その点で、小さなテストデータで測定される期待ロスＥＬは、近似値または推定値であると言える。

【０１４６】

分散関数ｆ（ｎ，ＥＬ）に入力する期待ロスＥＬの精度を上げるため、機械学習装置１００は、期待ロスＥＬの推定とテストデータサイズｎの選択を２回繰り返してもよい。例えば、機械学習装置１００は、小さなテストデータで測定した期待ロスＥＬを分散関数ｆ（ｎ，ＥＬ）に入力し、以下で説明する方法でテストデータサイズｎを仮選択する。機械学習装置１００は、データ集合から当該仮選択したサイズのテストデータを抽出し、抽出したテストデータを用いて期待ロスＥＬを再測定する。機械学習装置１００は、再測定した期待ロスＥＬを分散関数ｆ（ｎ，ＥＬ）に入力し、以下で説明する方法でテストデータサイズｎを再選択し、これを最終的なテストデータサイズと決定する。

【０１４７】

期待ロスＥＬを分散関数ｆ（ｎ，ＥＬ）に入力して期待ロスＥＬを固定すると、分散関数は、テストデータサイズｎと分散の推定値とを１対１に対応付ける対応関係を表す。機械学習装置１００は、分散関数のテストデータサイズｎを変動させながら分散の推定値を評価することで、適切なテストデータサイズｎを決定する。

【０１４８】

テストデータサイズｎと分散の推定値との対応関係は、テストデータサイズｎの増加に応じて、分散の推定値が下限に漸近するように減少する非線形曲線に相当する。テストデータサイズｎが小さいうちは、テストデータサイズｎの単位増加量あたりの分散の推定値の減少量が大きい。テストデータサイズｎが大きいほど、テストデータサイズｎの単位増加量あたりの分散の推定値の減少量が小さくなる。予測性能の測定値の信頼性を維持しつつテストデータサイズｎを小さくするため、適切なテストデータサイズｎは、分散の推定値が許容できる程度に小さい範囲で、できる限り小さいサイズとする。

【０１４９】

例えば、機械学習装置１００は、効果指標としてｆ（ｎ，ＥＬ）／ｆ（２＊ｎ，ＥＬ）を算出する。この効果指標は、テストデータサイズｎを２倍にした場合の分散の減少率に相当し、分散減少効果の評価指標である。効果指標の値が大きいほど分散減少効果が大きいことを示し、効果指標の値が小さいほど分散減少効果が小さいことを示す。テストデータサイズｎと分散の推定値の関係から、ｎが大きいほど効果指標の値は小さくなる。

【０１５０】

機械学習装置１００は、小さいテストデータサイズｎで効果指標の値を算出し、閾値と比較する。閾値は、１．１などと予め決めておく。効果指標の値が閾値以上である場合、機械学習装置１００は、テストデータサイズｎを２倍にし、効果指標の値が閾値未満になるまで上記を繰り返す。効果指標の値が閾値未満になると、機械学習装置１００は、その時点のテストデータサイズｎを適切なテストデータサイズとして決定する。

【０１５１】

なお、上記の方法におけるテストデータサイズｎの増加速度である「２倍」や閾値の「１．１」は調整可能パラメータであり、ユーザがこれらのパラメータを変更することも可能である。また、分散関数ｆ（ｎ，ＥＬ）から適切なテストデータサイズｎを探索する他の方法として、例えば、機械学習装置１００は、テストデータサイズｎを無限大にした場合の分散の推定値の下限を算出する。そして、機械学習装置１００は、分散の推定値が下限の所定倍（例えば、１．１倍）になるようなテストデータサイズｎを選択する。

【０１５２】

このようにして機械学習装置１００によって決定されるテストデータサイズは、訓練データサイズの２分の１または４分の１をテストデータサイズとする慣習的方法と比べて、十分に小さいサイズとなる。例えば、訓練データサイズが１００万レコードである場合、慣習的方法では、テストデータサイズが５０万レコードまたは２５万レコードとなる。これに対して、第２の実施の形態の方法によれば、予測性能の測定値の分散を慣習的方法と同程度に維持しつつ、テストデータサイズを数万レコード程度に削減できる。よって、予測性能の測定値の信頼性を維持しつつ、テスト処理を高速化できる。

【０１５３】

なお、第２の実施の形態で決定される最終的なテストデータサイズｎは、慣習的方法よりも十分に小さい。そのため、テストデータを用いて期待ロスＥＬを算出することを１回または２回行っても、全体のテスト処理の負荷は慣習的方法よりも十分に小さくなる。

【０１５４】

次に、機械学習装置１００の機能および処理手順について説明する。

図６は、機械学習装置の機能例を示すブロック図である。

機械学習装置１００は、データ記憶部１２１、制御情報記憶部１２２、学習結果記憶部１２３、モデル生成部１２４、テスト実行部１２５、テストサイズ決定部１２６および機械学習制御部１２７を有する。データ記憶部１２１、制御情報記憶部１２２および学習結果記憶部１２３は、例えば、ＲＡＭ１０２またはＨＤＤ１０３の記憶領域を用いて実現さ

れる。モデル生成部１２４、テスト実行部１２５、テストサイズ決定部１２６および機械学習制御部１２７は、例えば、ＣＰＵ１０１が実行するプログラムを用いて実現される。

【０１５５】

データ記憶部１２１は、訓練データまたはテストデータに使用可能な多数のレコードを含むデータ集合を記憶する。各レコードは、説明変数の値と教師ラベルである目的変数の値とを含む。データ集合は、数百万レコードなどサイズの大きなものであってもよい。機械学習装置１００は、ユーザからデータ集合を受け付けてもよいし、他の情報処理装置からデータ集合を受信してもよいし、センサデバイスからデータ集合を収集してもよい。

【０１５６】

制御情報記憶部１２２は、訓練データを用いたモデルの学習やテストデータを用いたモデルの予測性能の測定の過程で生成される各種の制御情報を記憶する。制御情報には、分散関数の生成に用いられる誤差プロファイルや分散関数のパラメータが含まれる。

【０１５７】

学習結果記憶部１２３は、機械学習の結果を記憶する。機械学習の結果には、学習されたモデルおよび当該モデルの予測性能の測定値が含まれる。

モデル生成部１２４は、機械学習によりモデルを生成する。モデル生成部１２４は、機械学習制御部１２７から機械学習アルゴリズムの指定と訓練データを受け付ける。モデル生成部１２４は、指定された機械学習アルゴリズムに従って、訓練データに含まれるレコードを用いてモデルの係数を決定することでモデルを学習する。機械学習アルゴリズムには、回帰分析、サポートベクタマシン、ランダムフォレストなどが含まれる。モデル生成部１２４は、学習されたモデルを機械学習制御部１２７に提供する。

【０１５８】

テスト実行部１２５は、モデルのテストを行う。テスト実行部１２５は、機械学習制御部１２７からモデルとテストデータを受け付ける。テスト実行部１２５は、テストデータのレコードに含まれる説明変数の値をモデルに入力し、モデルに従って目的変数の予測値を算出する。テスト実行部１２５は、テストデータのレコードに含まれる目的変数の真値とモデルから算出された予測値とを比較して、誤差を算出する。そして、テスト実行部１２５は、誤差を列挙した誤差プロファイルを生成する。

【０１５９】

テスト実行部１２５は、誤差プロファイルを機械学習制御部１２７に提供する。または、テスト実行部１２５は、誤差プロファイルを予測性能または期待ロスに変換し、予測性能または期待ロスを機械学習制御部１２７に提供する。予測性能の指標には、正答率、適合率、平均二乗誤差、二乗平均平方根誤差などが含まれる。予測性能または期待ロスは、テストデータに含まれる複数のレコードに対応する誤差から算出することができる。機械学習制御部１２７に提供される情報は、機械学習制御部１２７の要求に応じて変わる。

【０１６０】

テストサイズ決定部１２６は、テストデータサイズを決定する。まず、テストサイズ決定部１２６は、機械学習制御部１２７から誤差プロファイルを受け付ける。この誤差プロファイルは、mセットの小さな訓練データを用いて学習されたm個のモデルに対して、小さなテストデータを用いて測定された誤差を列挙したものである。テストサイズ決定部１２６は、この誤差プロファイルを用いて、予測性能の測定値の分散を推定するための分散関数のパラメータを決定する。分散関数のパラメータには、尖度Ｋ１と不動点Ｃｏｒ１ｖと期待バイアスＥＢ２が含まれる。テストサイズ決定部１２６は、分散関数の式や分散関数のパラメータの決定方法を予め知っている。テストサイズ決定部１２６は、分散関数のパラメータを機械学習制御部１２７に提供する。

【０１６１】

また、テストサイズ決定部１２６は、機械学習制御部１２７から、先に算出した分散関数のパラメータと、対象のモデルの訓練データサイズに対応する期待ロスＥＬを受け付ける。テストサイズ決定部１２６は、分散関数ｆ（ｎ，ＥＬ）に期待ロスＥＬを代入し、テストデータサイズｎを変えながら分散の推定値を算出する。そして、テストサイズ決定部

１２６は、適切なテストデータサイズｎを決定して機械学習制御部１２７に提供する。例えば、テストサイズ決定部１２６は、分散の推定値から算出される効果指標の値が閾値以上である範囲で、最大のテストデータサイズｎを検出する。

【０１６２】

機械学習制御部１２７は、機械学習を制御する。まず、機械学習制御部１２７は、モデルの学習および予測性能の測定の対象とする機械学習アルゴリズムおよび訓練データサイズを特定する。対象の機械学習アルゴリズムおよび訓練データサイズは、ユーザから指定されてもよいし、所定の規則に従って機械学習制御部１２７が選択してもよい。

【０１６３】

次に、機械学習制御部１２７は、テストサイズ決定部１２６に分散関数のパラメータを決定させる。ただし、分散関数のパラメータの決定は、予測性能を測定する対象のモデルが学習された後に行うようにすることも可能である。

【０１６４】

分散関数のパラメータの決定では、機械学習制御部１２７は、ｍセットの小さな訓練データと１セットの小さなテストデータを、データ記憶部１２１に記憶されたデータ集合から抽出する。機械学習制御部１２７は、ｍセットの訓練データをモデル生成部１２４に提供し、ｍ個のモデルをモデル生成部１２４から取得する。機械学習制御部１２７は、ｍ個のモデルと１セットのテストデータをテスト実行部１２５に提供し、誤差プロファイルをテスト実行部１２５から取得する。そして、機械学習制御部１２７は、誤差プロファイルをテストサイズ決定部１２６に提供し、分散関数のパラメータをテストサイズ決定部１２６から取得し、制御情報として制御情報記憶部１２２に格納する。

【０１６５】

次に、機械学習制御部１２７は、モデル生成部１２４に対象のモデルを学習させる。機械学習制御部１２７は、先に特定したサイズの訓練データを、データ記憶部１２１に記憶されたデータ集合から抽出する。機械学習制御部１２７は、抽出した訓練データをモデル生成部１２４に提供し、学習されたモデルをテスト実行部１２５から取得する。機械学習制御部１２７は、モデルを学習結果記憶部１２３に格納する。

【０１６６】

次に、機械学習制御部１２７は、対象のモデルの予測性能を測定するための適切なテストデータサイズをテストサイズ決定部１２６に決定させる。まず、機械学習制御部１２７は、学習結果記憶部１２３に記憶されたモデルと、分散関数のパラメータの決定の際に使用した小さなテストデータとを、テスト実行部１２５に提供する。機械学習制御部１２７は、このために小さなテストデータを保存しておいてもよい。また、機械学習制御部１２７は、分散関数のパラメータの決定の際に使用したテストデータに代えて、同等のサイズのテストデータを、データ記憶部１２１に記憶されたデータ集合から抽出してもよい。

【０１６７】

機械学習制御部１２７は、テスト実行部１２５から期待ロスを取得し、制御情報記憶部１２２に記憶された分散関数のパラメータと期待ロスをテストサイズ決定部１２６に提供する。ただし、機械学習制御部１２７は、対象のモデルを用いて期待ロスを測定する代わりに、回帰分析などの統計的方法によって期待ロスを推定してもよい。機械学習制御部１２７は、テストサイズ決定部１２６からテストデータサイズを取得する。

【０１６８】

すると、機械学習制御部１２７は、データ記憶部１２１に記憶されたデータ集合から、決定されたサイズのテストデータを抽出する。テストデータに含まれるレコードは訓練データと重複しないことが好ましい。機械学習制御部１２７は、抽出したテストデータと学習結果記憶部１２３に記憶されたモデルとをテスト実行部１２５に提供する。機械学習制御部１２７は、テスト実行部１２５から予測性能の測定値を取得し、学習結果記憶部１２３に格納する。ただし、機械学習制御部１２７は、上記のテストデータに対して、更新された期待ロスをテスト実行部１２５から取得し、更新された期待ロスに基づいて、更新されたテストデータサイズをテストサイズ決定部１２６から取得してもよい。

【０１６９】
モデルの学習と予測性能の測定が完了すると、機械学習制御部１２７は、モデルおよび予測性能の測定値を出力する。例えば、機械学習制御部１２７は、表示装置１１１にモデルおよび予測性能の測定値を表示する。機械学習制御部１２７は、他の出力デバイスにモデルおよび予測性能の測定値を出力してもよい。また、例えば、機械学習制御部１２７は、他の情報処理装置にモデルおよび予測性能の測定値を送信する。
【０１７０】
図７は、誤差プロファイルテーブルの例を示す図である。
誤差プロファイルテーブル１３１は、制御情報記憶部１２２に記憶される。誤差プロファイルテーブル１３１は、ｍセットの訓練データとｎレコードのテストデータとの間で網羅的に算出されたｍ×ｎ個の誤差を記憶する。誤差プロファイルテーブル１３１の列は、訓練データD_1，D_2，…，D_mに対応する。誤差プロファイルテーブル１３１の行は、テストデータのｎ個のレコードに含まれる入力値X_1，X_2，…，X_nに対応する。１つの訓練データD_kから学習された１つのモデルに、テストデータの１つのレコードに含まれる入力値X_iを入力することで、予測値と真値との差である誤差e_{ik}が算出される。
【０１７１】
図８は、分散関数テーブルの例を示す図である。
分散関数テーブル１３２は、制御情報記憶部１２２に記憶される。分散関数テーブル１３２は、尖度Ｋ１、不動点Ｃｏｒ１ｖおよび期待バイアスＥＢ２の３つのパラメータに対応する値を記憶する。これら３つのパラメータは、数式（４４）に含まれるパラメータであって、訓練データサイズに依存しないパラメータである。分散関数テーブル１３２に記憶される値は、誤差プロファイルテーブル１３１から算出される。
【０１７２】
図９は、機械学習の手順例を示すフローチャートである。
（Ｓ１０）機械学習制御部１２７は、機械学習アルゴリズムと訓練データサイズを指定する。機械学習アルゴリズムと訓練データサイズの指定はユーザから受け付けてもよい。
【０１７３】
（Ｓ１１）機械学習制御部１２７は、データ記憶部１２１からｍセットの小サイズの訓練データと１セットの小サイズのテストデータを抽出する。例えば、１万レコードの訓練データが１０セット抽出され、５０００レコードのテストデータが１セット抽出される。
【０１７４】
（Ｓ１２）モデル生成部１２４は、ステップＳ１０で指定された機械学習アルゴリズムに従って、ｍセットの訓練データからｍ個のモデルを学習する。
（Ｓ１３）テスト実行部１２５は、ステップＳ１２で学習されたｍ個のモデルに、ステップＳ１１のテストデータの各レコードを入力して誤差を算出し、算出した誤差を列挙した誤差プロファイルテーブル１３１を生成する。具体的には、テスト実行部１２５は、１つのモデルとテストデータの１つのレコードの組毎に、レコードに含まれる説明変数の値をモデルに入力し、モデルから算出された目的変数の予測値とレコードに含まれる真値との差を誤差として算出する。例えば、１０個のモデルと５０００レコードのテストデータから、１０×５０００個の誤差を含む誤差プロファイルテーブル１３１が生成される。
【０１７５】
（Ｓ１４）テストサイズ決定部１２６は、誤差プロファイルテーブル１３１から、所定の数式に従って、分散関数ｆ（ｎ，ＥＬ）を規定するパラメータの値を決定する。パラメータには、尖度Ｋ１と不動点Ｃｏｒ１ｖと期待バイアスＥＢ２が含まれる。ここで決定されるパラメータの値は、使用するデータ集合と指定された機械学習アルゴリズムに依存するものである一方、訓練データサイズに依存しないものである。
【０１７６】
（Ｓ１５）機械学習制御部１２７は、データ記憶部１２１から、ステップＳ１０で指定されたサイズの訓練データを抽出する。
（Ｓ１６）モデル生成部１２４は、ステップＳ１０で指定された機械学習アルゴリズム

に従って、ステップＳ１５で抽出された訓練データからモデルを学習する。
【０１７７】
図１０は、機械学習の手順例を示すフローチャート（続き）である。
（Ｓ１７）テスト実行部１２５は、ステップＳ１６で学習されたモデルに、ステップＳ１１で抽出された小サイズのテストデータの各レコードを入力して誤差を算出する。ただし、ステップＳ１１で抽出されたものとは異なるテストデータを使用してもよい。
【０１７８】
（Ｓ１８）テスト実行部１２５は、ステップＳ１７で算出された誤差から、所定の数式に従って、ステップＳ１６で学習されたモデルの期待ロスＥＬを推定する。
（Ｓ１９）テストサイズ決定部１２６は、ステップＳ１４で決定されたパラメータの値をもつ分散関数ｆ（ｎ，ＥＬ）に、ステップＳ１８で推定された期待ロスＥＬを代入する。テストサイズ決定部１２６は、分散関数ｆ（ｎ，ＥＬ）により算出される分散が所定条件を満たす範囲で、最大のテストデータサイズｎ１を判定する。例えば、テストサイズ決定部１２６は、テストデータサイズｎを２倍にした場合の分散の減少率を示す効果指標の値と所定の閾値とを比較し、効果指標の値が閾値未満になるまでテストデータサイズｎを２倍にすることを繰り返す。これにより、最大のテストデータサイズｎ１が選択される。
【０１７９】
（Ｓ２０）機械学習制御部１２７は、データ記憶部１２１から、ステップＳ１９で判定されたサイズｎ１のテストデータを抽出する。
（Ｓ２１）テスト実行部１２５は、ステップＳ１６で学習されたモデルに、ステップＳ２０で抽出されたテストデータの各レコードを入力して誤差を算出する。
【０１８０】
（Ｓ２２）テスト実行部１２５は、ステップＳ２１で算出された誤差から、所定の数式に従って、ステップＳ１６で学習されたモデルの期待ロスＥＬを再推定する。
（Ｓ２３）テストサイズ決定部１２６は、ステップＳ１４で決定されたパラメータの値をもつ分散関数ｆ（ｎ，ＥＬ）に、ステップＳ２２で再推定された期待ロスＥＬを代入する。テストサイズ決定部１２６は、分散関数ｆ（ｎ，ＥＬ）により算出される分散が所定条件を満たす範囲で、最大のテストデータサイズｎ２を判定する。テストデータサイズｎ２の判定方法は、ステップＳ１９と同様の方法でよい。
【０１８１】
（Ｓ２４）機械学習制御部１２７は、データ記憶部１２１から、ステップＳ２３で判定されたサイズｎ２のテストデータを抽出する。
（Ｓ２５）テスト実行部１２５は、ステップＳ１６で学習されたモデルに、ステップＳ２４で抽出されたテストデータの各レコードを入力して誤差を算出する。テスト実行部１２５は、算出された誤差から、当該モデルの予測性能の測定値を算出する。
【０１８２】
（Ｓ２６）機械学習制御部１２７は、ステップＳ１６で学習されたモデルとステップＳ２５で算出された予測性能の測定値を、学習結果記憶部１２３に保存する。また、機械学習制御部１２７は、モデルおよび予測性能の測定値を表示装置１１１に表示する。
【０１８３】
なお、上記のフローチャートでは、対象となるモデルの期待ロスＥＬの推定を２回繰り返している。期待ロスＥＬの推定を１回だけ行う場合、上記のステップＳ１９～Ｓ２２を省略することができる。また、対象となるモデルを使用せずに統計的方法により期待ロスＥＬを推定する場合、上記のステップＳ１７～Ｓ２２を省略することができる。
【０１８４】
第２の実施の形態の機械学習装置１００によれば、複数セットの小さい訓練データと１セットの小さいテストデータを用いて、同一のデータ集合および機械学習アルゴリズムのもとで生じる誤差の分布を示す誤差プロファイルが生成される。誤差プロファイルに基づいて、期待ロスとテストデータサイズを引数としてもち、予測性能の測定値の分散を算出する分散関数が決定される。そして、大きい訓練データを用いて学習された対象モデルの

期待ロスが推定され、分散関数が示すテストデータサイズと分散の対応関係に基づいて、対象モデルの予測性能を測定するための適切なテストデータサイズが決定される。

【０１８５】

テストデータサイズは、予測性能の測定値が実用上十分な信頼性をもつ範囲、すなわち、その分散が許容できる範囲で、できる限り小さいサイズに決定される。これにより、テストデータサイズが小さ過ぎることにより予測性能の測定値の信頼性が低下することを抑制できる。また、テストデータサイズが大き過ぎることにより予測性能の測定値の信頼性向上に寄与しない無駄なテスト処理が発生することを抑制でき、テスト処理の負荷を軽減してテスト時間を短縮できる。よって、学習されたモデルの予測性能を高信頼かつ短時間で測定することができ、テスト処理を効率化することができる。例えば、テストデータサイズを訓練データサイズの２分の１から４分の１程度とする慣習的方法と比べて、測定値の分散を同程度に抑えつつ、テストデータサイズを削減することができる。

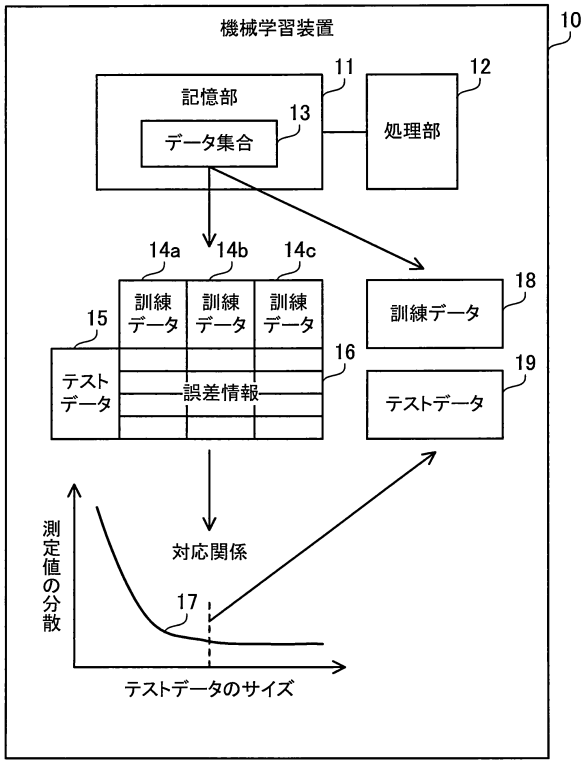
【符号の説明】

【０１８６】

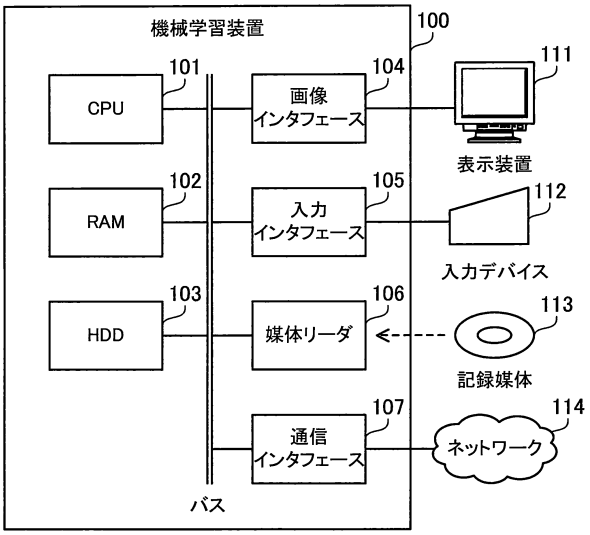
１０　機械学習装置
１１　記憶部
１２　処理部
１３　データ集合
１４ａ，１４ｂ，１４ｃ，１８　訓練データ
１５，１９　テストデータ
１６　誤差情報
１７　対応関係

【図面】

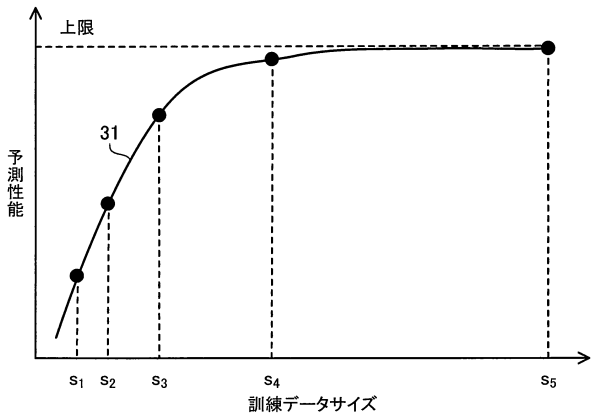
【図１】


機械学習装置
10
記憶部
11
データ集合
13
処理部
12
14a
14b
14c
訓練データ
訓練データ
訓練データ
15
テストデータ
誤差情報
16
訓練データ
18
テストデータ
19
対応関係
測定値の分散
17
テストデータのサイズ

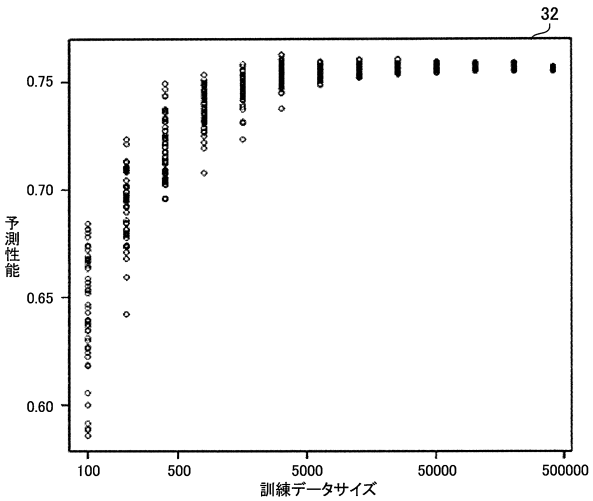
【図２】


機械学習装置
100
CPU
101
RAM
102
HDD
103
画像インタフェース
104
入力インタフェース
105
媒体リーダ
106
通信インタフェース
107
バス
111
表示装置
112
入力デバイス
113
記録媒体
114
ネットワーク

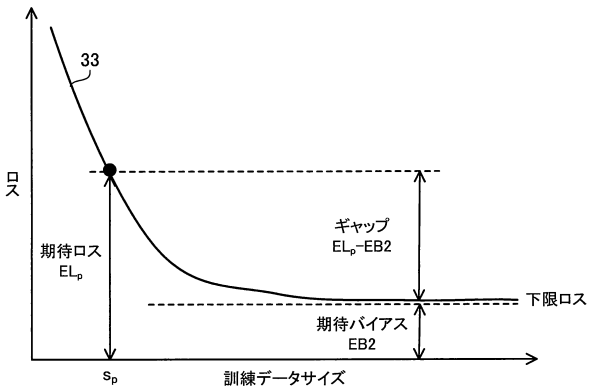
【図３】


上限
31
予測性能
s1
s2
s3
s4
s5
訓練データサイズ

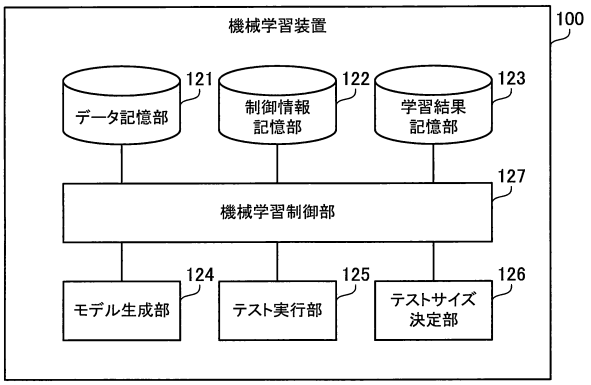
【図４】


32
0.75
0.70
0.65
0.60
予測性能
100
500
5000
50000
500000
訓練データサイズ

【図５】



【図６】



【図７】

誤差プロファイルテーブル　131

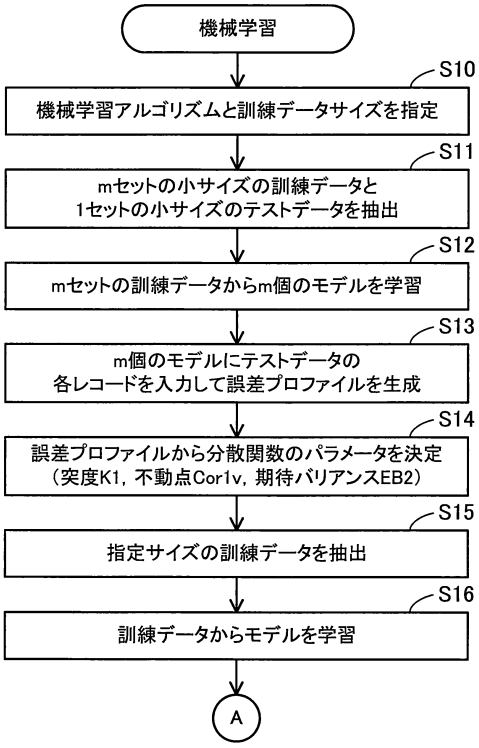
		訓練データ					
		D_1	D_2	・・・	D_k	・・・	D_m
テストデータ	X_1	e_{11}	e_{12}	・・・	e_{1k}	・・・	e_{1m}
	X_2	e_{21}	e_{22}	・・・	e_{2k}	・・・	e_{2m}
	⋮	⋮	⋮		⋮		⋮
	X_i	e_{i1}	e_{i2}	・・・	e_{ik}	・・・	e_{im}
	⋮	⋮	⋮		⋮		⋮
	X_n	e_{n1}	e_{n2}	・・・	e_{nk}	・・・	e_{nm}

【図８】

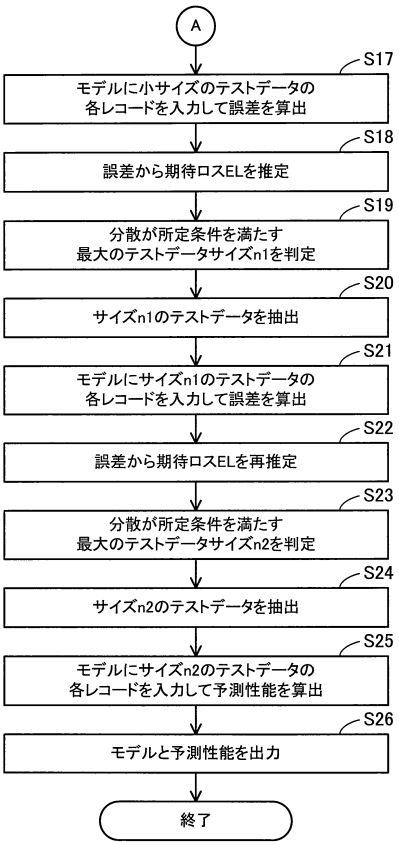
分散関数テーブル　132

パラメータ	値
尖度 K1	・・・
不動点 Cor1v	・・・
期待バイアス EB2	・・・

【図９】


機械学習
S10
機械学習アルゴリズムと訓練データサイズを指定
S11
mセットの小サイズの訓練データと
1セットの小サイズのテストデータを抽出
S12
mセットの訓練データからm個のモデルを学習
S13
m個のモデルにテストデータの
各レコードを入力して誤差プロファイルを生成
S14
誤差プロファイルから分散関数のパラメータを決定
（突度K1，不動点Cor1v，期待バリアンスEB2）
S15
指定サイズの訓練データを抽出
S16
訓練データからモデルを学習
A

【図１０】


A
S17
モデルに小サイズのテストデータの
各レコードを入力して誤差を算出
S18
誤差から期待ロスELを推定
S19
分散が所定条件を満たす
最大のテストデータサイズn1を判定
S20
サイズn1のテストデータを抽出
S21
モデルにサイズn1のテストデータの
各レコードを入力して誤差を算出
S22
誤差から期待ロスELを再推定
S23
分散が所定条件を満たす
最大のテストデータサイズn2を判定
S24
サイズn2のテストデータを抽出
S25
モデルにサイズn2のテストデータの
各レコードを入力して予測性能を算出
S26
モデルと予測性能を出力
終了

フロントページの続き

(56)参考文献 特表２０１５－５２５４１３（ＪＰ，Ａ）
特開平９－５４７６４（ＪＰ，Ａ）
特開２０１９－１１３９１５（ＪＰ，Ａ）
特開２０１７－４９６７４（ＪＰ，Ａ）

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｆ ８／００－８／３８
８／６０－８／７７
９／４４－９／４４５
９／４５１
Ｇ０６Ｎ ３／００－３／１２
７／０８－９９／００