(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5398682号
(P5398682)
(45)発行日　平成26年1月29日(2014.1.29)　(24)登録日　平成25年11月1日(2013.11.1)

(51)Int.Cl.　　F I
G06T 7/00 (2006.01)　　G06T 7/00 350B
G06N 3/00 (2006.01)　　G06N 3/00 560A

請求項の数 5　外国語出願　(全 9 頁)

(21)出願番号　特願2010-242078 (P2010-242078)
(22)出願日　平成22年10月28日 (2010.10.28)
(65)公開番号　特開2011-118883 (P2011-118883A)
(43)公開日　平成23年6月16日 (2011.6.16)
審査請求日　平成25年8月27日 (2013.8.27)
(31)優先権主張番号　12/631,590
(32)優先日　平成21年12月4日 (2009.12.4)
(33)優先権主張国　米国 (US)

早期審査対象出願

(73)特許権者　597067574
ミツビシ・エレクトリック・リサーチ・ラボラトリーズ・インコーポレイテッド
アメリカ合衆国、マサチューセッツ州、ケンブリッジ、ブロードウエイ　201
201 BROADWAY, CAMBRIDGE, MASSACHUSETTS 02139, U.S.A.
(74)代理人　100110423
弁理士　曾我　道治
(74)代理人　100094695
弁理士　鈴木　憲七
(74)代理人　100111648
弁理士　梶並　順

最終頁に続く

(54)【発明の名称】局所的学習のためのトレーニング点の近傍を選択するための方法

(57)【特許請求の範囲】
【請求項１】
局所的学習のためのトレーニング点のセットから、或るクエリポイントの近くのトレーニング点のサブセットを選択するための方法であって、
トレーニング点のセットＸ、クエリポイントx_qを与えるステップと、
以下の式に従って、トレーニング点のセットＸから、前記クエリポイントx_qの近くのトレーニング点のサブセットX_Nを求めるステップとを含み、
【数１】

$$\mathrm{argmax}_{X \subset \mathcal{X}} G(X) = D_T(X) + \lambda e^{-H(X)}$$

ただし、関数ａｒｇｍａｘは関数Ｇを最大にするＸの値を返し、ｐ＝１，２の場合の
【数２】

$$D_T(X) = \sum_{x \in X} \exp(-||x - x_q||_p)$$

は前記クエリポイントから前記トレーニング点のサブセットX_nへの累積類似度であり、$e^{-H(X)}$はＸによって誘導される分布の逆範囲を求め、λは０より大きな制御パラメータであり、Ｈはシャノンエントロピーであり、
前記与えるステップ及び前記求めるステップはプロセッサにおいて実行される、選択するための方法。
【請求項２】

以下の式
【数３】

$$H(\mathcal{N}(\mu,\Sigma)) = \frac{1}{2}\log|\Sigma| + \frac{C}{2}(\log 2\pi e)$$

に従って、前記シャノンエントロピーを推定するステップをさらに含み、ただし、μは前記サブセットX_N内の前記トレーニング点の平均であり、Σは前記サブセットX_N内の前記トレーニング点の共分散であり、Cは前記トレーニング点の前記セットの次元数である、請求項１に記載の方法。
【請求項３】
前記トレーニング点の前記サブセットX_Nを用いて分類法をトレーニングするステップをさらに含む、請求項１に記載の方法。
【請求項４】
前記トレーニング点の前記サブセットX_Nを用いて回帰法をトレーニングするステップをさらに含む、請求項１に記載の方法。
【請求項５】
前記回帰法として異分散サポートベクトル回帰を用いるステップをさらに含む、請求項４に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は包括的には局所的な教師あり学習に関し、より詳細には、単一のクエリポイントに基づいて、トレーニングデータセットからトレーニング点のサブセットを選択することに関する。
【背景技術】
【０００２】
教師あり学習では、クエリポイントから成る入力ベクトルに基づいて未知の出力を予測する関数を推定するために、トレーニングデータが用いられる。局所的学習法では、所与のクエリポイントに対して、そのクエリポイントの「近くにある」トレーニング点によって関数が求められる。近さは、或る距離メトリックによって判断することができる。
【０００３】
局所的学習法の例は、最近傍回帰及び分類、並びに局所重み付け回帰を含む。２つの応用例として、過去の値に基づいて、或る時系列の未来の値を予測すること、及びピクセル値に基づいて、或る画像内に或る特定の物体が存在するか否かを検出することが挙げられる。
【０００４】
そのような問題において、トレーニングデータセットDは複数のペアから成るセットであり、
【０００５】
【数１】

$$D = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_M}, y_M)\} \subset \mathcal{X} \times \mathcal{R},$$

【０００６】
である。ただし、Xは入力パターンを表し、たとえば、$X = R^d$である。各ペアは入力ベクトルx_i及び出力y_iを含む。関数
【０００７】
【数２】

$$\hat{y} = F(\mathbf{x})$$

【０００８】

は対応する入力ベクトルから出力を推定し、その関数はトレーニングデータセットから学習される。

【０００９】

局所的学習法において、クエリポイントX_q毎に、入力クエリポイントX_qの近くにある、トレーニングセット内のトレーニングデータ点だけに基づいて、局所関数Ｆ(ｘ)が学習される。クエリポイントの近くにあるトレーニング点は通常、距離メトリックに従って、トレーニングデータセット内のｋ個の最も近い点から選択される。代替的には、選択されるトレーニング点は、クエリポイントからの或る距離しきい値ｄ未満である。

【００１０】

局所的学習法の背景にある概念は、データは入力空間の異なる部分において異なる特性を有することができること、及びクエリポイントの近くにあるデータは、所与の入力から所望の出力を予測する関数を学習するのに最も有用なはずであるということである。

【００１１】

１つの応用例において、毎日の電力需要を予測することが望まれる。１年のうちの異なる時点では、異なる要因が需要負荷に影響を及ぼす可能性がある。クエリポイントが夏日に対応する場合には、トレーニングデータセット内の夏日だけに基づいて、関数Ｆ()を学習することが好都合である可能性がある。

【００１２】

しかしながら、ｋ個の最近傍点、又は或る距離ｄ内にある全ての近傍点を用いることが、必ずしも最良の性能を与えるとは限らない。

【発明の概要】

【発明が解決しようとする課題】

【００１３】

いずれのトレーニング点が近傍に属するかを求めるための方法と共に、局所近傍の新たな概念を提供することが望まれる。

【課題を解決するための手段】

【００１４】

方法は、累積類似度を最大にすることによって、トレーニング点のセットから、１つのクエリポイントに近いトレーニング点のサブセットを選択し、その累積類似度は、クエリポイントとサブセット内の各点との類似度、及びサブセット内の点の互いの類似度を評価する。

【図面の簡単な説明】

【００１５】

【図１Ａ】トレーニングデータ、及びこの発明の実施の形態による方法によって選択されるトレーニングデータのサブセットのプロット図である。

【図１Ｂ】トレーニングデータ、及び従来技術の方法によって選択されるトレーニングデータのサブセットのプロット図である。

【図２Ａ】トレーニングデータ、及びこの発明の実施の形態による方法によって選択されるトレーニングデータのサブセットのプロット図である。

【図２Ｂ】トレーニングデータ、及び従来技術の方法によって選択されるトレーニングデータのサブセットのプロット図である。

【図３】この発明の実施の形態による、トレーニングデータを選択するための方法のフローチャートである。

【発明を実施するための形態】

【００１６】

図３に示されるように、この発明の実施の形態は、単一の入力クエリポイントx_q３０５の近くにある、複数の点から成る局所近傍X_N３０２を選択するための方法を提供する。近傍点の全てがコンパクトであることが望ましい。この明細書において定義されるように、複数の点が或る距離メトリックに従って互いに相対的に近い場合には、そのセットはコンパクトである。この方法のステップは、当該技術分野において既知のように、メモリ

及び入力／出力インターフェースを備えるプロセッサにおいて実行することができる。

【００１７】

従来の最近傍法は、結果として得られる近傍がコンパクトであるか否かを考慮することなく、クエリポイントの近くにある近傍点を選択する。

【００１８】

この発明の実施の形態による方法は、入力トレーニングデータ(３０１参照)が入力空間内で不均一に分布するときに性能を改善するために、コンパクトであることの判定基準を含む。

【００１９】

この発明による局所近傍点のサブセットX_N３０２は(３１０参照)以下の通りである。

【００２０】

【数３】

$$X_N = \mathrm{argmax}_{X \subset \mathcal{X}} G(X) = D_T(X) + \lambda e^{-H(X)}, \lambda > 0 \qquad (1)$$

【００２１】

ただし、関数ａｒｇｍａｘ３１０は関数Ｇを最大にするパラメータＸの値を返し、ｐ＝１，２の場合に、

【００２２】

【数４】

$$D_T(X) = \sum_{x \in X} \exp(-||x - x_q||_p)$$

【００２３】

はクエリポイント３０５からトレーニングサブセット

【００２４】

【数５】

$$X \subset \mathcal{X}$$

【００２５】

への累積類似度であり、$e^{-H(X)}$はＸによって誘導される分布の逆範囲(inverse range)を求め、λは制御パラメータである。分布の「範囲」によって、この発明では、サンプル空間が離散的である場合には、サンプル空間内の点の数を意味し、サンプル空間が指数関数的なエントロピーの特性に従う実線である場合には、確率密度関数が０でない区間の長さを意味する。Ｈはシャノンエントロピーである。シャノンエントロピーは、確率変数の値が未知であるときに失う平均情報量の指標である。

【００２６】

この発明では、以下のようにガウス分布を仮定してシャノンエントロピーを推定する。

【００２７】

【数６】

$$H(\mathcal{N}(\mu, \Sigma)) = \frac{1}{2} \log |\Sigma| + \frac{C}{2} (\log 2\pi e) \qquad (2)$$

【００２８】

ただし、μはサブセット内の点の平均であり、Σはその共分散である。Ｃは入力トレーニング点３０１の次元数である。

【００２９】

式(１)の目的は、或るクエリポイントの近くの最も大きなクラスタへの累積類似度を最大にすることである。この発明の目的は、トレーニングデータパターンの基本的な構造に適応できるような方法で、トレーニングデータのサブセット３０２を見つけることである。

【００３０】

この問題の組み合わせ最適化特性は、従来の最近傍法において用いられる貪欲法との重要な相違点である。この発明において最大にする、式(１)において定義される目的関数は、優モジュラ性として知られている数学的特性を有する。全てのｘ，ｙ∈Ｒ^{k}に対して、

【００３１】

【数７】

$$f(x \vee y) + f(x \wedge y) \geq f(x) + f(y)$$

【００３２】

である場合には、関数

【００３３】

ｆ：Ｒ^{k}→Ｒ

【００３４】

は優モジュラ性である。ただし、

【００３５】

【数８】

$$x \vee y$$

【００３６】

はｘ及びｙの成分毎の最大値を表し、

【００３７】

【数９】

$$x \wedge y$$

【００３８】

は成分毎の最小値を表す。

【００３９】

優モジュラ関数を最大にすることは、劣モジュラ関数を最小にすることと同じである。それゆえ、この発明では、この関数を最適化するために、劣モジュラ性最適化の従来の手段を適用することができる。

【００４０】

上記の手順を用いて、複数の点から成る最適なサブセットが求められた後に、任意の分類又は回帰法をトレーニングするために(３２０)、その点のサブセットを用いることができる。

【００４１】

図１Ａ，１Ｂ，２Ａ，２Ｂを用いて、２つの単純な例を説明し、ｋ＝１５の場合の、この発明の劣モジュラモデル選択法と、従来のｋ最近傍点選択法とを比較する。それらの図は、単一のクエリポイント１０１と、サブセット１０２とを示す。これらの図において、時間間隔［－８，８］内に１００個の合成入力データ点ｘ$_{i}$が均一な間隔で配置されており、対象となる出力データ点ｙ$_{i}$＝２ｓｉｎ(２πｘ)が、ｘ＝０における０．５からｘ＝１における１．５まで線形に増加する標準偏差を有するガウス雑音で汚染される。

【００４２】

これらの図は、この発明の方法が従来の方法よりも優れていることを明示する。劣モジュラ法では、近傍点が適応的に選択され、ここで図１Ａ及び図１Ｂにおいて示されるように分布の先端では点の数が少なく、図２Ａ及び図２Ｂにおいて示されるように、末端に近いほど数が多くなり、それは、この発明の例においてガウス雑音の汚染が増加することに一致する。

【００４３】

異分散サポートベクトル回帰

上記のように選択される近傍トレーニングデータを用いて、任意の回帰又は分類法をトレーニングすることができる。ここで、異分散サポートベクトル回帰のための１つのその

ような技法を説明する。異分散サポートベクトル回帰は、サポートベクトル回帰を拡張したものであり、局所近傍を用いて局所回帰関数を見つける。統計学において、一連の確率変数が異なる分散を有するときに、それらの確率変数は異分散である。

【００４４】

異分散サポートベクトル回帰は、形式

【００４５】

$$F(x) = w^T x + b$$

【００４６】

の関数Ｆ(ｘ)を推定する。ただし、w^Tは転置演算子Ｔを伴うベクトルであり、ｂはスカラーである。

【００４７】

その関数は、以下の最適化問題を解くことによって求められる。

【００４８】

【数１０】

$$\min_{\mathbf{w},b,\epsilon,\xi,\xi^*} \frac{1}{2N}\sum_{i=1}^{N}\mathbf{w}^T(2N\mathrm{I}+\Sigma_i)\mathbf{w} + C\sum_{i=1}^{N}(\xi_i+\xi_i^*+\epsilon)^p$$
$$\text{s.t.} y_i - (\mathbf{w}^T\mathbf{x_i}+b) \le \epsilon+\xi_i^*$$
$$(\mathbf{w}^T\mathbf{x_i}+b) - y_i \le \epsilon+\xi_i \qquad (3)$$
$$\xi_i^*, \xi_i \ge 0, \forall i,$$

【００４９】

ただし、ＩはＮ×Ｎ恒等行列であり、Ｎは入力ベクトルの次元数であり、ξ_i及びξ^*_iはスラック変数であり、εは誤差許容範囲であり、$p \in \{1, 2\}$はペナルティタイプを決定し、そして

【００５０】

【数１１】

$$\Sigma_i = \frac{1}{k_i+1}\sum_{\mathbf{x}\in X_i}(\mathbf{x}-\bar{\mathbf{x}}_\mathbf{i})(\mathbf{x}-\bar{\mathbf{x}}_\mathbf{i})^T$$

【００５１】

はx_iの近傍内のトレーニング点のための実験的な共分散である。ただし、X_iは近傍点のサブセットであり、k_iはX_i内の点の数であり、

【００５２】

【数１２】

$$\bar{\mathbf{x}}_\mathbf{i}$$

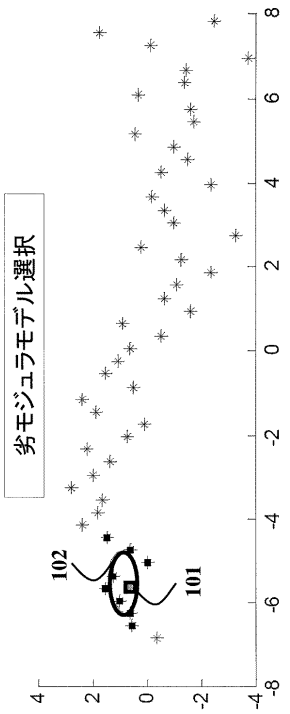
【００５３】

はこれらの近傍点の平均である。複数の点から成るこの近傍は、上記の劣モジュラ性最適化技法を用いて選択することができる。

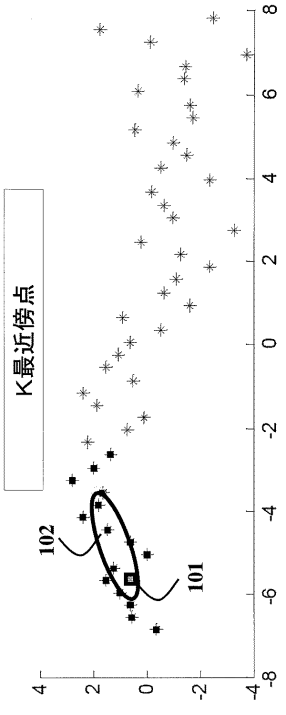
【００５４】

この発明を、好ましい実施の形態の例として説明してきたが、この発明の精神及び範囲内で他のさまざまな適合及び変更を行えることが理解されるべきである。したがって、この発明の真の精神及び範囲内に入るすべての変形及び変更を包含することが、添付の特許請求の範囲の目的である。

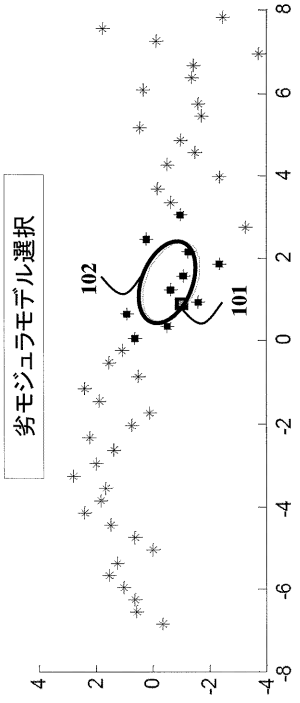
【図１Ａ】



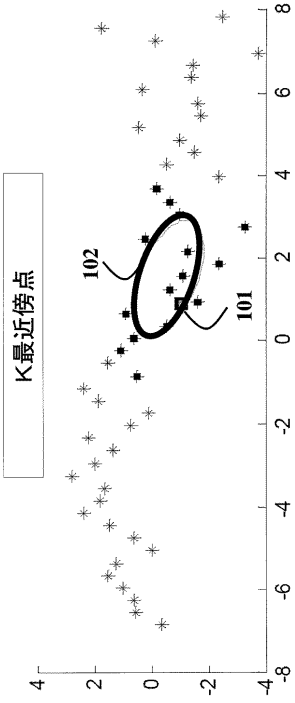
【図１Ｂ】



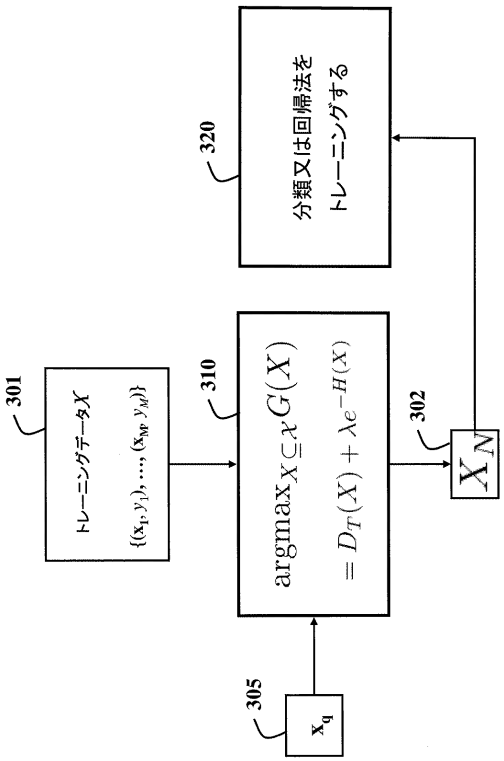
【図２Ａ】



【図２Ｂ】



【図３】


320
分類又は回帰法を
トレーニングする
301
トレーニングデータX
{(x1, y1), ..., (xM, yM)}
310
argmax$_{X \subseteq \mathcal{X}} G(X)$
$= D_T(X) + \lambda e^{-H(X)}$
302
X_N
305
x_q

フロントページの続き

(74)代理人 100122437
弁理士 大宅 一宏
(74)代理人 100147566
弁理士 上田 俊一
(74)代理人 100161171
弁理士 吉田 潤一郎
(74)代理人 100161115
弁理士 飯野 智史
(72)発明者 ケヴィン・ダブリュ・ウィルソン
アメリカ合衆国、マサチューセッツ州、ケンブリッジ、ウィンター・ストリート ８０、ユニット １
(72)発明者 シャオハン・ジアン
アメリカ合衆国、マサチューセッツ州、ケンブリッジ、ヴァッサー・ストリート ３２

審査官 佐藤 実

(56)参考文献 特開平８－１０６２９５（ＪＰ，Ａ）
特開２００９－２５９１０９（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ ７／００
Ｇ０６Ｎ ３／００