

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2012-123782

(P2012-123782A)

(43) 公開日 平成24年6月28日(2012.6.28)

(51) Int.Cl.

G06N 5/04 (2006.01)

F1

G06N 5/04 580A

テーマコード (参考)

審査請求 未請求 請求項の数 7 O L 外国語出願 (全 17 頁)

(21) 出願番号 特願2011-230987 (P2011-230987)
 (22) 出願日 平成23年10月20日 (2011.10.20)
 (31) 優先権主張番号 12/961,895
 (32) 優先日 平成22年12月7日 (2010.12.7)
 (33) 優先権主張国 米国 (US)

(71) 出願人 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、ブロードウェイ 201
 201 BROADWAY, CAMBR
 IDGE, 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

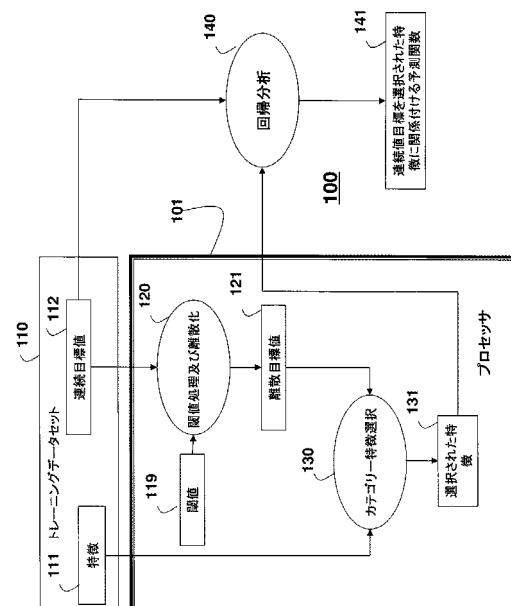
(54) 【発明の名称】 連続値回帰分析において用いられる特徴を選択する方法

(57) 【要約】 (修正有)

【課題】連続値回帰分析問題等の連続値問題のための特徴を選択する方法を提供する。

【解決手段】方法100へのトレーニングデータ110の入力は特徴及び対応する目標値を含み、目標値は連続しており、特徴ごとに1つの目標値が存在する。各閾値は閾値に対して閾値処理及び離散化され、離散化された目標値が生成される。次に、離散目標値を用いて特徴にカテゴリー特徴選択が適用され、選択された特徴がもたらされる。選択された値は任意の回帰分析において用いることができる。

【選択図】図1



【特許請求の範囲】**【請求項 1】**

連続値回帰分析において用いられる特徴を選択する方法であって、
入力としてトレーニングデータセットを提供するステップであって、該トレーニングデータセットは特徴及び対応する目標値を含み、該目標値は連続しており、前記特徴ごとに 1 つの目標値が存在するものと、
閾値に対して各前記目標値を閾値処理及び離散化するステップであって、離散化された目標値を生成するものと、
前記離散化された目標値を用いて前記特徴にカテゴリー特徴選択を適用するステップであって、選択された特徴をもたらすものと、
を含み、前記ステップはプロセッサにおいて実行される、方法。

10

【請求項 2】

前記選択された特徴を用いて連続値回帰分析を実行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 3】

前記閾値処理及び離散化するステップは、前記目標値が分割される 1 つ又は複数の閾値に対するものである、請求項 1 に記載の方法。

【請求項 4】

単一の閾値を、前記トレーニングデータセットにおける前記目標値の中央値になるように選択するステップであって、結果として平衡の取れたクラスをもたらすものをさらに含む、請求項 3 に記載の方法。

20

【請求項 5】

前記 1 つ又は複数の閾値は用途に特有の知識に基づく、請求項 2 に記載の方法。

【請求項 6】

前記回帰分析は非線形である、請求項 1 に記載の方法。

【請求項 7】

前記回帰分析は最小絶対偏差及び l_1 正則化を用いた異分散サポートベクター回帰である、請求項 1 に記載の方法。

【発明の詳細な説明】

30

【技術分野】**【0001】**

この発明は、包括的には連続値回帰分析(continuous-valued regression analysis)に関し、より詳細には、連続値回帰分析を実行するときに利用可能な特徴のセットから特徴のサブセットを選択することに関する。

【背景技術】**【0002】**

回帰分析は、従属変数と 1 つ又は複数の独立変数との間の関係をモデリングする。回帰分析は、独立変数のうちの任意の 1 つが変化し、一方でその他の独立変数が同じままであるとき、通常の従属変数がどのように変化するかを求めることができる。最も一般的には、回帰分析は、独立変数を所与として従属変数の条件付き期待値を推定する。

40

【0003】

この発明が特に対象とするのは、連続値回帰分析において用いられる特徴を選択することである。回帰分析のための手順は、ニューラルネットワーク及びサポートベクターマシン(SVM)を含む。回帰分析の通常の用途には、時系列予測、たとえば過去の値に基づいた電力需要の未来値の予測、及び利用可能な測定値に基づく対象となる未知の量の予測、たとえば身長、体重、血圧、及び髪の高さの測定値に基づくたとえば人間の寿命の予測が含まれる。

【0004】

特徴選択は、回帰分析において用いられる利用可能な特徴のサブセットを求める。寿命を予測する用途の上記の例では、有用な特徴のサブセットは、身長、体重、及び血圧を含

50

むことができる一方、髪の長さは有用でない。この用途では、特徴選択手順は、有用な特徴、たとえば身長、体重、及び血圧のサブセットのみを選択すべきであり、無用な特徴、たとえば髪の高さを除外すべきである。無用な特徴を除去することによって、特徴選択は後続の予測のための時間を低減することができる。無用な特徴を除去することによって、特徴選択は後続の予測の正確度を改善し、解釈がより容易なモデルをもたらすこともできる。

【発明の概要】

【発明が解決しようとする課題】

【0005】

多くの特徴選択手順は、相関等の一次従属の単純な測度を用いて有用な特徴を選択する。これらの手法は、変数間の関係が非線形であるとき、失敗する可能性がある。ラッパ法は、特定の潜在的に非線形の回帰分析問題を評価することによって一度に少数の特徴を貪欲に選択する。ラッパ法は、特徴の小さなサブセットを貪欲に選択するので、特徴の最良な全体的な組み合わせを求めることができない。ラッパ法は多くの場合に計算集約的であり、回帰方法をサブルーチンとして直接組み込むので、その特定の回帰分析方法に直結する。

10

【0006】

既知のRELIEF特徴選択手順によって、他の特徴選択方法の望ましくない特性のほとんどが回避される。参照によりこの明細書に援用される、2007年6月19日にLee他に発行された米国特許第7,233,931号「Feature regulation for hierarchical decision learning」を全体的に参照されたい。この方法は貪欲でなく、計算集約的でなく、特定の回帰分析方法に結びついていない。しかしながら、RELIEF手順は分類問題及びカテゴリー問題、すなわち従属変数が離散値の小さなセットからの値をとることができる問題の場合にのみ機能する。カテゴリー問題の例は、病気検出問題であり、この問題において、従属変数は病気の存否を示す2つの可能な値のうちの1つをとることができる。カテゴリー問題と対照的に連続値問題は値の無限集合からの値、たとえば全ての実数をとることができる。この場合、従属変数によってとられる値を「目標値」と呼ぶ。

20

【課題を解決するための手段】

【0007】

この発明の実施の形態は、連続値回帰分析問題等の連続値問題のための特徴を選択する方法を提供する。この方法は貪欲でなく、特徴と目標値との間の非線形関係が存在する問題に適用することができ、この方法は任意の回帰分析方法と共に用いることができる。

30

【図面の簡単な説明】

【0008】

【図1】この発明の実施の形態による連続値回帰分析において用いられる特徴を選択する方法の流れ図である。

【発明を実施するための形態】

【0009】

この発明の実施の形態は、連続値回帰分析において用いられる特徴を選択するための方法を提供する。二項分類のための特徴選択は既知であり、特に既知であるのはRELIEF法である。上記の米国特許第7,233,931号を参照されたい。RELIEF法は、多数の無用な特徴が存在するときに良好に機能する。RELIEF法の基本的な着想は、局所トレーニングを用いて任意の非線形問題を局所線形問題のセットに分解し、次に1(エル)₁正則化を用いて大きなマージンフレームワークにおいて大域的に特徴関連性を推定することである。

40

【0010】

RELIEF法は、通例同じクラスにある近傍の値を有する特徴を選択する。近傍の値を有する例が必ずしも同じクラスを共有しない特徴は選択されない。RELIEF法は同じクラス又は異なるクラスからの近傍の特徴を検査するので、当然二項分類に適用される。

50

【 0 0 1 1 】

しかしながら、R E L I E F 法は連続値回帰分析に適用されない。したがって、この発明の目的は、R E L I E F 法を連続値回帰分析に適応させることである。

【 0 0 1 2 】

R E L I E F 法を連続値回帰分析に適応させる 1 つの以前の試みは、特徴値に対する目標値の変化率を局所的に推定し、特徴値における変化が目標値の変化に対して予測可能な関係を有する特徴を選択する。米国特許第 7 , 2 3 3 , 9 3 1 号を参照されたい。この手法はいくつかの状況において機能するが、特徴と目標値との間の関係を推定するために選択される近傍のサイズの影響を非常に受けやすい可能性がある。

【 0 0 1 3 】

図 1 に示し以下でより詳細に説明するこの発明の方法では、連続出力をカテゴリ - 離散値出力 1 2 1 にカテゴリ化する。この方法は、連続値目標値 y を、1 つ又は複数の閾値 y_m において閾値処理する (1 2 0)。閾値は目標値が分割される 2 つ以上のクラスを定義する。

【 0 0 1 4 】

単一の閾値 y_m をトレーニングデータセット 1 1 0 内の目標値の中央値となるように選択することによって、結果としてクラスの平衡が取られる、すなわち各クラス内の例の数が実質的に同じになる。用途に特有の知識を用いて他の閾値を選択することができる。

【 0 0 1 5 】

単一の閾値の場合、第 1 のクラス及び第 2 のクラスを以下のように定義する。

【 0 0 1 6 】

【 数 1 】

$$y' = \begin{cases} 1 & y < y_m \text{ の場合} \\ 2 & y \geq y_m \text{ の場合} \end{cases}$$

【 0 0 1 7 】

ここで、 y' は離散クラスラベルである。

【 0 0 1 8 】

y_m を、トレーニングセットにわたる目標値 y の中央値となるように選択することは、これによって第 1 のクラスの例の数が第 2 のクラスの例の数と実質的に同じになることが保証されるので、通例良好な閾値となる。

【 0 0 1 9 】

閾値処理を実行して二項クラスを割り当てた後、ロジスティック R E L I E F 法又は任意の他のカテゴリ - 特徴選択方法においてクラス値 1 2 1 を特徴値 1 1 1 と共に用いることができる。

【 0 0 2 0 】

非線形回帰分析方法

特徴選択を実行した後、選択された特徴を回帰分析方法 1 4 0 の入力として用いることができる。この発明の特徴選択方法は、予測値に対して非線形関係を有する特徴を選択することができる。したがって、当然ながら選択された特徴に非線形回帰分析方法を用いる。

【 0 0 2 1 】

回帰分析

この発明の回帰分析は、最小絶対偏差及び l_1 正則化を用いた異分散 (異なる分散を有する) サポートベクター回帰であり、トレーニングデータセット D 1 1 0 は M 個の対のセット

【 0 0 2 2 】

10

20

30

40

【数 2】

$$\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_M, Y_M)\} \subset \mathcal{X} \times \mathcal{R}$$

【0023】

を含み、ここで \mathcal{X} は入力パターンの空間、たとえば $\mathcal{X} = \mathbb{R}^d$ を示す。

【0024】

各対は、入力ベクトル \mathbf{x}_i 、及び目標出力値 y_i を含む。関数

【0025】

【数 3】

$$\hat{Y} = F(\mathbf{X})$$

10

【0026】

は、対応する入力ベクトル \mathbf{x}_i から目標出力 y_i を推定する。目標出力はトレーニングデータセット 110 から求められる。

【0027】

この発明の回帰分析は、以下の最適化問題を解くことによってトレーニングされる。

【0028】

【数 4】

$$\min_{\mathbf{W}, \beta, b, \sigma, \xi, \xi^*} \lambda \|\mathbf{W}\|_1 + \frac{1}{2} \beta^T (\mathbf{K} + \Delta) \beta + C \sum_i (\xi_i + \xi_i^* + \epsilon_i)$$

20

$$s.t. \quad Y_i - (\mathbf{W}^T \mathbf{X}_i + K_i + b) \leq \epsilon_i + \xi_i^*,$$

$$(\mathbf{W}^T \mathbf{X}_i + \beta^T K_i + b) - Y_i \leq \epsilon_i + \xi_i,$$

$$= \quad \xi_i^*, \xi_i, \epsilon_i \geq 0, \forall i.$$

【0029】

ここで、 ξ_i 及び ξ_i^* はスラック変数、すなわち誤分類の度合いを示す変数であり、 ϵ_i は誤差許容度であり、 λ は L_1 正則化の強度を制御するパラメーター、すなわちトレーニング誤差の最小化と比較した \mathbf{W} の L_1 ノルムの最小化の相対的重要度であり、 \mathbf{W} は線形係数ベクトルであり、 \mathbf{K} はカーネル行列であり、 β はカーネル行列のための係数ベクトルであり、 C は適合度を制御するパラメーターである。

30

【0030】

新たな試験例のためのラベルが予測関数に従って推定される。

【0031】

【数 5】

$$\phi''(\mathbf{X}) = \mathbf{W}^T \mathbf{X} + \sum_i \beta_i K(\mathbf{X}_i, \mathbf{X}) + b$$

40

【0032】

予測関数における線形項は、高度に相関した特徴、すなわち従属変数と強力な線形関係にある特徴がトレーニングデータ内に存在する場合に有用である。この発明の回帰分析における非線形カーネル項の特性は、特徴と従属変数との間の非線形関係を利用することによる補完的役割を有する。

【0033】

連続値回帰のための特徴選択方法

この発明の方法は以下を用いる。

1. カテゴリーデータにも適用することができる特徴選択方法。これを「カテゴリー特徴選択方法」と呼ぶ。この発明の実施態様において、カテゴリーデータのためのこの発明

50

の特徴選択方法として R E L I E F の一変形を用いる。

２．回帰分析方法。この発明の実施態様では、最小絶対偏差及び l_1 正則化を用いた異分散サポートベクター回帰を用いる。

【 0 0 3 4 】

図 1 は、連続値回帰分析のための特徴を選択するためのこの発明の方法 1 0 0 を示している。この方法は、当該技術分野において既知のメモリ及び入力／出力インターフェースを含むプロセッサ 1 0 1 において実行することができる。

【 0 0 3 5 】

この方法への入力はトレーニングデータ 1 1 0 である。トレーニングデータは、特徴 1 1 1 と、対応する連続目標値 y 1 1 2 とを含む。

10

【 0 0 3 6 】

連続値目標値 y は、１つ又は複数の閾値 y_m 1 1 9 において閾値処理及び離散化され (1 2 0)、離散目標値 1 2 1 が生成される。閾値は目標値を分割することができる 2 つ以上のクラスを定義する。

【 0 0 3 7 】

閾値のうちの 1 つが、結果として 2 つの平衡が取られたクラスをもたらすための目標値の中央値として選択される。用途に特有の知識に基づいて他の閾値を選択することができる。

【 0 0 3 8 】

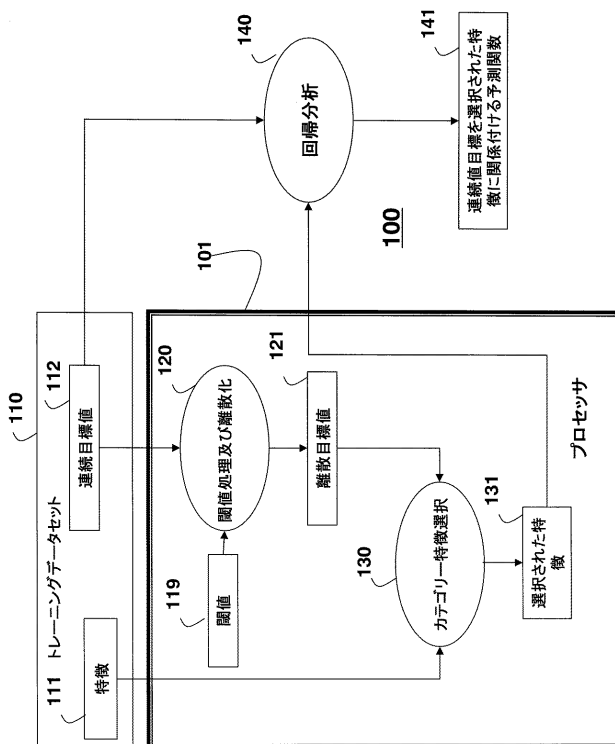
離散値目標 1 2 1 を用いてカテゴリー特徴選択 1 3 0 が特徴 1 1 1 に適用され、選択された特徴 1 3 1 がもたらされる。

20

【 0 0 3 9 】

選択された特徴 1 3 1 及び連続目標値 1 1 2 が回帰分析方法 1 4 0 に入力され、連続値目標を選択された特徴に関係付ける予測関数 1 4 1 が生成される。

【 図 1 】



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、パラモ・ストリート 19

(72)発明者 ユボ・チェン

アメリカ合衆国、ニュージャージー州、ジャージー・シティ、リバー・ドライブ・サウス 45、
アパートメント 603

【 外国語明細書 】

Title of the Invention

Method for Selecting Features Used in Continuous-Valued Regression Analysis

Field of the Invention

This invention relates generally to continuous-valued regression analysis, and more particularly to selecting a subset of features from a set of available features when performing continuous-valued regression analysis.

Background of the Invention

Regression analysis models a relationship between a dependent variable and one or more independent variables. Regression analysis can determine how a typical dependent variable changes when any one of the independent variables is varied, while the other independent variables remain the same. Most commonly, regression analysis estimates a conditional expectation of the dependent variable given the independent variables.

Of particular interest to this invention is selecting features used in continuous-valued regression analysis. Procedures for regression analysis include neural networks, and support vector machines (SVM). Typical applications of regression analysis include time series prediction, e.g., the prediction of future values of an electrical power demand based on past values, and prediction of an unknown quantity of interest based on available measurements, e.g., the prediction of a person's lifespan based on measurements of height, weight, blood pressure, and hair length, for example.

Feature selection determines a subset of the available features used in regression analysis. In the above example of an application for predicting lifespan, the subset of useful features can include height, weight, and blood pressure, while hair length is not useful. In this application, the feature selection procedure should only select the subset of useful features, e.g. height, weight, and blood pressure, and the procedure should exclude the useless feature, e.g., hair length. By eliminating useless features, the feature selection can reduce the time for subsequent prediction. By eliminating useless features, feature selection can also improve the accuracy of subsequent predictions, and lead to models that are easier to interpret.

Many feature selection procedures use simple measures of linear dependence, such as correlation, to select useful features. Those approaches can fail when the relationships among the variables are nonlinear. Wrapper techniques greedily select a small number of features at a time by evaluating a specific, potentially nonlinear, regression analysis problem. Because wrapper techniques greedily select a small subset of the features, wrapper techniques cannot determine the best overall combination of features. Wrapper techniques are often computationally intensive, and because wrapper techniques directly incorporate a regression method as a subroutine, wrapper techniques are directly tied to that particular regression analysis method.

The well known RELIEF feature selection procedure avoids most of the undesirable properties of other feature selection methods, see generally U.S. Patent 7,233,931 issued to Lee, et al. on June 19, 2007, "Feature regulation for hierarchical decision learning," incorporated herein by reference. That method is not greedy, not computationally intensive, and not tied to a specific regression analysis method.

However, the RELIEF procedure works only for classification and categorical problems, i.e., problems in which the dependent variable can take a value from a small set of discrete values. An example of a categorical problem would be a disease detection problem, where the dependent variable can take one of two possible values indicating presence or absence of the disease. In contrast to categorical problems, *continuous-valued* problems have dependent variables that can take values from an infinite set of values, for example all real numbers. In this case, we refer to the values taken on by the dependent variable as “*target values*.”

Summary of the Invention

The embodiments of the invention provide a method for selecting features for continuous-valued problems, such as continuous-valued regression analysis problems. The method is not greedy, can be applied to problems in which nonlinear relationships between features and target values exist, and the method can be used with any regression analysis method.

Brief Description of the Drawings

Fig. 1 is a flow diagram of a method for selecting features used in continuous-valued regression analysis according to embodiments of the invention.

Detailed Description of the Preferred Embodiment

The embodiments of our invention provide a method for selecting features used in continuous-valued regression analysis. Feature selection for binary classification are well known, particularly the well known RELIEF method, see U.S. Patent

7,233,931 above. The RELIEF method works well when there are a large number of useless features. The basic idea of the RELIEF method is to decompose an arbitrary nonlinear problem into a set of locally linear problems using local training, and then estimate the feature relevance globally in a large margin framework with the ℓ_1 regularization.

The RELIEF method selects features that have nearby values that are usually in the same class. Features for which examples with nearby values do not necessarily share the same class are not selected. Because the RELIEF method examines neighboring features from the same class, or from different classes, it naturally applies to binary classification.

However, the RELIEF method does not apply to continuous-valued regression analysis. Therefore, an object of the invention is to adapt the RELIEF method to continuous-valued regression analysis.

One previous attempt to adapt the RELIEF method to continuous-valued regression analysis locally estimates a rate of change of a target value with respect to the feature values, and selects features for which changes in the feature value has a predictable relationship to the change in target value, see U.S. Patent 7,233,931. That approach works in some situations, but it can be very sensitive to a size of the neighborhood selected for estimating relationships between features and target values.

In our method as shown in Fig. 1 and described in greater detail below, we categorize the continuous output into categorical discrete-valued output 121. Our

method thresholds 120 the continuous-valued target values y , at one or more threshold values y_m . The threshold values define two or more classes into which the target values are partitioned.

Selecting a single threshold value y_m to be a median of the target values in a training data set 110 results in balanced classes, i.e., the number of examples in each class are substantially the same. Other threshold values can be selected based on application-specific knowledge.

For the case of a single threshold value, we define a first class and a second class as

$$y' = \begin{cases} 1 & \text{if } y < y_m \\ 2 & \text{if } y \geq y_m \end{cases},$$

where y' is a discrete class label.

Selecting y_m to be the median value of target values y over the training set is usually a good threshold value because this guarantees that the number of examples of the first class is substantially the same as the number of examples of the second class.

After performing the thresholding to assign a binary class, the class values 121 can be used along with the feature values 111 in the logistic RELIEF method, or in any other categorical feature selection method.

Nonlinear Regression Analysis Method

After feature selection has been performed, the selected features can be used as input to a regression analysis method 140. Our feature selection method can select features with nonlinear relationships to the predicted value. Therefore, it is natural to use a nonlinear regression analysis method with the selected features.

Regression Analysis

Our regression analysis is a heteroscedastic (having differing variances) support vector regression with least absolute deviation and ℓ_1 regularization, the set of training data \mathcal{D} 110 includes a set of M pairs

$$\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_M, Y_M)\} \subset \mathcal{X} \times \mathcal{R},$$

where X denotes a space of input patterns, e.g., $X = R^d$.

Each pair includes an input vector \mathbf{X}_i , and a target output value y_i . A function $\hat{Y} = F(\mathbf{X})$ estimates the target output y_i from the corresponding input vector \mathbf{X}_i . The target output is determined from the training data set 110.

Our regression analysis is trained by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{W}, \beta, b, \sigma, \xi, \xi^*} \quad & \lambda \|\mathbf{W}\|_1 + \frac{1}{2} \beta^T (\mathbf{K} + \Delta) \beta + C \sum_i (\xi_i + \xi_i^* + \epsilon_i) \\ \text{s.t.} \quad & Y_i - (\mathbf{W}^T \mathbf{X}_i + \beta^T K_i + b) \leq \epsilon_i + \xi_i^*, \\ & (\mathbf{W}^T \mathbf{X}_i + \beta^T K_i + b) - Y_i \leq \epsilon_i + \xi_i, \\ = \quad & \xi_i^*, \xi_i, \epsilon_i \geq 0, \forall i. \end{aligned}$$

where ξ_i and ξ_i^* are slack variables, i.e. the variables that indicate the degree of misclassification, ε is an error tolerance, λ is a parameter that controls a strength of the ℓ_1 regularization, i.e., the relative importance of minimizing the ℓ_1 norm of \mathbf{W} compared to minimizing the training error, \mathbf{W} is a linear coefficient vector, \mathbf{K} is a kernel matrix, β is a coefficient vector for the kernel matrix, and C is a parameter that controls a goodness of fit.

The label for a new test example is estimated according to a prediction function

$$\phi''(\mathbf{X}) = \mathbf{W}^T \mathbf{X} + \sum_i \beta_i K(\mathbf{X}_i, \mathbf{X}) + b$$

The linear term in the prediction function is helpful when highly correlated features, i.e., features with a strong linear relationship with the dependent variable, exist in the training data. The nature of nonlinear kernel term in our regression analysis has complementary role by exploiting nonlinear relationships between the features and the dependent variable.

Method for Feature Selection for Continuous-Values Regression

Our method uses:

1. A feature selection method that can also be applied to categorical data. We refer to this as the “categorical feature selection method.” In our implementation, we use a variant of RELIEF as our feature selection method for categorical data.
2. A regression analysis method. In our implementation, we use heteroscedastic support vector regression with least absolute deviation and ℓ_1 regularization.

Figure 1 shows our method 100 for selecting features for continuous-valued regression analysis. The method can be performed in a processor 101 including memory and input/output interfaces as known in the art.

Input to the method is training data 110. The training data includes features 111 and corresponding continuous target values y 112.

The continuous-valued target values y are thresholded and discretized 120 at one or more threshold values y_m 119 to produce discrete target values 121. The threshold values define two or more classes into which the target values can be partitioned.

One of the threshold values is selected as the median of the target values to result in two balanced classes. Other thresholds can be selected based on application-specific knowledge.

Categorical feature selection 130 is applied to the features 111 using the discrete-values targets 121 to produce selected features 131.

The selected features 131 and the continuous target values 112 are input into the regression analysis method 140 to produce a prediction function 141 relating the continuous-values targets to the selected features.

1. A method for selecting features used in continuous-valued regression analysis, comprising the steps of:

providing a training data set as input, wherein the training data set includes features and corresponding target values, wherein the target values are continuous, and there is one target value for each feature;

thresholding and discretizing each target value with respect to a threshold value to produce a discretized target value; and

applying categorical feature selection to the features using the discretized target values to produce selected features, wherein the steps are performed in a processor.

2. The method of claim 1, further comprising the step of:

performing continuous-valued regression analysis using the selected features.

3. The method of claim 1, wherein the thresholding and discretizing step is with respect to one or more threshold values into which the target values are partitioned.

4. The method of claim 3, further comprising the step of:

selecting a single threshold value to be a median of the target values in the training data set to result in balanced classes.

5. The method of claim 2, wherein the one or more threshold values are based on application-specific knowledge.

6. The method of claim 1, wherein the regression analysis is nonlinear.

7. The method of claim 1, wherein the regression analysis is a heteroscedastic support vector regression with least absolute deviation and ℓ_1 regularization.

Abstract

A method selects features used in continuous-valued regression analysis. Training data input to the method includes features and corresponding target values, wherein the target values are continuous, and there is one target value for each feature. Each threshold value is thresholded and discretized with respect to a threshold value to produce a discretized target value. Then, categorical feature selection is applied to the features, using the discrete target values, to produce selected features. The selected values can be used in any regression analysis.

Representative Drawing

Figure 1

