

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号

特許第7068085号

(P7068085)

(45)発行日 令和4年5月16日(2022.5.16)

(24)登録日 令和4年5月6日(2022.5.6)

(51)国際特許分類

F I

G 0 6 F 16/20 (2019.01)

G 0 6 F 16/20

G 0 6 F 16/28 (2019.01)

G 0 6 F 16/28

G 0 6 F 17/15 (2006.01)

G 0 6 F 17/15

G 0 6 F 17/18 (2006.01)

G 0 6 F 17/18

Z

請求項の数 8 (全20頁)

(21)出願番号 特願2018-137210(P2018-137210)

(22)出願日 平成30年7月20日(2018.7.20)

(65)公開番号 特開2020-13511(P2020-13511A)

(43)公開日 令和2年1月23日(2020.1.23)

審査請求日 令和2年11月10日(2020.11.10)

(73)特許権者 000005108

株式会社日立製作所

東京都千代田区丸の内一丁目6番6号

(74)代理人 110002365

特許業務法人サンネクスト国際特許事務

所

(72)発明者 平山 淳一

東京都千代田区丸の内一丁目6番6号

株式会社日立製作所内

(72)発明者 芹田 進

東京都千代田区丸の内一丁目6番6号

株式会社日立製作所内

(72)発明者 工藤 文也

東京都千代田区丸の内一丁目6番6号

株式会社日立製作所内

最終頁に続く

(54)【発明の名称】 特徴量生成装置および特徴量生成方法

(57)【特許請求の範囲】

【請求項1】

目的変数に係る複数の説明変数において組合せを取って特徴量を生成する際、前記目的変数との相関係数がしきい値以上になる特徴量を生成する特徴量生成部を備え、

前記特徴量生成部は、

前記複数の説明変数から生成し得る複数の特徴量の各々について、一の特徴量と前記一の特徴量とは異なる任意の他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が前記しきい値以上になるか否かを判定し、前記複数の特徴量から前記しきい値未満の特徴量を除去し、除去して残った特徴量同士の組合せを取った特徴量を新たな特徴量として生成する、

ことを特徴とする特徴量生成装置。

【請求項2】

前記特徴量生成部は、説明変数の組合せ数が N (N 2の整数)の特徴量を生成する際、説明変数の組合せ数が k ($1 \leq k \leq N$ の整数)の複数の特徴量の各々について、一の特徴量と前記一の特徴量とは異なる任意の他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が前記しきい値以上になるか否かを判定し、前記複数の特徴量から前記しきい値未満の特徴量を除去し、除去して残った説明変数の組合せ数が「1」の特徴量と除去して残った組合せ数が k の特徴量との組合せを取った特徴量を説明変数の組合せ数が $k+1$ の新たな特徴量として生成する処理を、説明変数の組合せ数が N の特徴量を生成するまで k が1から順に繰り返す、

ことを特徴とする請求項 1 に記載の特徴量生成装置。

【請求項 3】

前記特徴量生成部は、前記複数の特徴量から前記しきい値未満の特徴量を除去する際、前記目的変数の値と前記一の特徴量の値とを示す点群に基づいて、前記一の特徴量と前記他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が十分に大きくなるように、前記目的変数の値を変更することなく前記一の特徴量の値を変更することで前記点群を移動し、移動した点群と前記目的変数との相関係数である相関見込み値を計算し、計算した相関見込み値が前記しきい値以上であるか否かを判定する、ことを特徴とする請求項 1 に記載の特徴量生成装置。

【請求項 4】

前記特徴量生成部は、前記目的変数の値と前記一の特徴量の値とを示す点群に基づいて、前記一の特徴量と前記他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が十分に大きくなるための前記点群の移動パターンを示す移動境界直線を計算し、計算した移動境界直線に基づいて前記点群を移動する、ことを特徴とする請求項 3 に記載の特徴量生成装置。

【請求項 5】

目的変数に係る複数の説明変数において組合せを取って特徴量を生成する際、前記目的変数との相関係数がしきい値以上になる特徴量を生成する特徴量生成方法であって、特徴量生成部が、前記複数の説明変数から生成し得る複数の特徴量の各々について、一の特徴量と前記一の特徴量とは異なる任意の他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が前記しきい値以上になるか否かを判定し、前記複数の特徴量から前記しきい値未満の特徴量を除去し、除去して残った特徴量同士の組合せを取った特徴量を新たな特徴量として生成する、ことを特徴とする特徴量生成方法。

【請求項 6】

前記特徴量生成部が、説明変数の組合せ数が N ($N \geq 2$ の整数) の特徴量を生成する際、説明変数の組合せ数が k ($1 \leq k \leq N$ の整数) の複数の特徴量の各々について、一の特徴量と前記一の特徴量とは異なる任意の他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が前記しきい値以上になるか否かを判定し、前記複数の特徴量から前記しきい値未満の特徴量を除去し、除去して残った説明変数の組合せ数が「1」の特徴量と除去して残った組合せ数が k の特徴量との組合せを取った特徴量を説明変数の組合せ数が $k + 1$ の新たな特徴量として生成する処理を、説明変数の組合せ数が N の特徴量を生成するまで k が 1 から順に繰り返す、ことを特徴とする請求項 5 に記載の特徴量生成方法。

【請求項 7】

前記特徴量生成部が、前記複数の特徴量から前記しきい値未満の特徴量を除去する際、前記目的変数の値と前記一の特徴量の値とを示す点群に基づいて、前記一の特徴量と前記他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が十分に大きくなるように、前記目的変数の値を変更することなく前記一の特徴量の値を変更することで前記点群を移動し、移動した点群と前記目的変数との相関係数である相関見込み値を計算し、計算した相関見込み値が前記しきい値以上であるか否かを判定する、ことを特徴とする請求項 5 に記載の特徴量生成方法。

【請求項 8】

前記特徴量生成部が、前記目的変数の値と前記一の特徴量の値とを示す点群に基づいて、前記一の特徴量と前記他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が十分に大きくなるための前記点群の移動パターンを示す移動境界直線を計算し、計算した移動境界直線に基づいて前記点群を移動する、ことを特徴とする請求項 7 に記載の特徴量生成方法。

【発明の詳細な説明】

10

20

30

40

50

【技術分野】

【0001】

本発明は特徴量生成装置および特徴量生成方法に関し、例えば目的変数との相関係数がしきい値以上になる特徴量を生成する特徴量生成装置および特徴量生成方法に適用して好適なものである。

【背景技術】

【0002】

近年、ビッグデータと呼ばれる大量のデータを分析し、今まで人が勘と経験で行ってきた意思決定を支援するシステムの開発が急速に発展してきている。これらのデータ分析は、テーブル形式のデータに対して行われることが多い。例えば、ある目的変数を変動させる他の説明変数が何であるかを見つけ出すための相関分析、説明変数群の値から目的変数の値を予測する回帰分析といった機械学習・統計分析を主な手法としている。また、説明変数になんらかの加工を施した変数を新たな説明変数とすることもある。

10

【0003】

近年、あるイベントの発生有無を示す説明変数に対し、全ての説明変数の組合せを網羅的に計算した特徴量群を生成し、その中から目的変数に寄与する特徴量を抽出する技術が開示されている（特許文献1参照）。

【先行技術文献】

【特許文献】

【0004】

【文献】特開2017-102937号公報

20

【発明の概要】

【発明が解決しようとする課題】

【0005】

特許文献1に記載の特徴量の計算においては、入力テーブルとして入力された全ての説明変数に対して取り得る全ての組合せを網羅的に計算している。そのため、説明変数の数が多くなるにつれて、計算される特徴量の数が指数関数的に増大（組合せ爆発）する。

【0006】

説明変数の数 = n 、組合せ数 = r とした場合、計算される特徴量の数は、 $nCr + nC(r-1) + nC(r-2) + \dots + nC1$ となる。例えば、 $n = 100$ 、 $r = 3$ の場合、特徴量の数は、166750個となり、計算時間が膨大になる可能性がある。

30

【0007】

このような計算時間の膨大は、データ分析者にとって、分析サイクル時間の増大につながっている。また、入力テーブルに入力する説明変数の数を意図的に少なくするような事前作業により、組合せ爆発を回避するなどの作業工数のため、分析コストの増大につながっている。更に、組合せ爆発により、計算機に必要なメモリが膨大になり、計算コストの増大につながっている。

【0008】

本発明は以上の点を考慮してなされたもので、目的変数に対する相関の大きい特徴量を生成するための計算量を削減し得る特徴量生成装置等を提案しようとするものである。

40

【課題を解決するための手段】

【0009】

かかる課題を解決するため本発明においては、例えば、目的変数に係る複数の説明変数において組合せを取って特徴量を生成する際、前記目的変数との相関係数がしきい値以上になる特徴量を生成する特徴量生成部を備え、前記特徴量生成部は、前記複数の説明変数から生成し得る複数の特徴量の各々について、一の特徴量と前記一の特徴量とは異なる任意の他の特徴量との組合せを取ったときの特徴量と前記目的変数との間で取り得る相関係数が前記しきい値以上になるか否かを判定し、前記複数の特徴量から前記しきい値未満の特徴量を除去し、除去して残った特徴量同士の組合せを取った特徴量を新たな特徴量として生成するようにした。

50

【 0 0 1 0 】

上記構成では、複数の特徴量からしきい値未満の特徴量を除去し、除去して残った特徴量同士の組合せを取った特徴量を新たな特徴量として生成するので、目的変数に対する相関の大きい特徴量を生成するための計算量を削減することができる。

【 発明の効果 】

【 0 0 1 1 】

本発明によれば、目的変数に対する相関の大きい特徴量を適切に生成することができる。

【 図面の簡単な説明 】

【 0 0 1 2 】

【 図 1 】 第 1 の実施の形態による特徴量選択システムに係る構成の一例を示す図である。

10

【 図 2 】 第 1 の実施の形態による特徴量抽出を説明するためのデータの一例を示す図である。

【 図 3 】 第 1 の実施の形態による特徴量抽出を説明するためのデータの一例を示す図である。

【 図 4 】 第 1 の実施の形態による特徴量抽出を説明するためのデータの一例を示す図である。

【 図 5 】 第 1 の実施の形態による特徴量選択システムの全体フローを説明するための図である。

【 図 6 】 第 1 の実施の形態による特徴量生成部の処理の流れを説明するための模式図である。

20

【 図 7 】 第 1 の実施の形態による特徴量生成部が実行する処理に係るフローチャートの一例を示す図である。

【 図 8 】 第 1 の実施の形態による散布図の一例を示す図である。

【 図 9 】 第 1 の実施の形態による散布図の一例を示す図である。

【 図 1 0 】 第 1 の実施の形態による散布図の一例を示す図である。

【 図 1 1 】 第 1 の実施の形態による相関がなるべく大きくなるときの移動先の点群の考え方について補足するための図である。

【 図 1 2 】 第 1 の実施の形態による相関見込み値の計算処理に係るフローチャートの一例を示す図である。

【 図 1 3 】 第 1 の実施の形態による移動境界直線の計算処理に係るフローチャートの一例を示す図である。

30

【 図 1 4 】 第 1 の実施の形態による移動境界直線のパラメータの定め方の例を示す図である。

【 発明を実施するための形態 】

【 0 0 1 3 】

以下図面について、本発明の一実施の形態を詳述する。

【 0 0 1 4 】

本実施の形態に示す技術は、データ分析技術およびデータマイニング技術に関するものである。例えば、テーブル形式で蓄積されたデータに対して、目的変数に関連のある説明変数を抽出する技術に関する。特に、相関分析を用いることで、ユーザが指定した目的変数に対する相関の大きい特徴量（説明変数の組合せ）を抽出する特徴量抽出に関する。

40

【 0 0 1 5 】

本実施の形態によれば、例えば、説明変数同士の組合せからなる特徴量を計算する際に、より少ない時間で、目的変数に対する相関の高い特徴量を抽出することができる。また、本実施の形態によれば、例えば、目的変数に対する相関の高い特徴量を抽出する際に、より少ない特徴量の計算回数で、同様の抽出結果を得ることができる。

【 0 0 1 6 】

以下では、便宜上その必要があるときは、複数のセクションまたは実施の形態に分割して説明するが、特に明示した場合を除き、それらは互いに無関係なものではなく、一方は他方の一部または全部の変形例、詳細、補足説明等の関係にある。また、以下では、要素の

50

数等（個数、数値、量、範囲等を含む。）に言及する場合、特に明示した場合、原理的に明らかに特定の数に限定される場合等を除き、その特定の数に限定されるものではなく、特定の数以上でも以下でもよい。

【 0 0 1 7 】

（ 1 ）第 1 の実施の形態

図 1 において、1 は全体として第 1 の実施の形態による特徴量選択システムを示す。

【 0 0 1 8 】

特徴量選択システム 1 は、目的変数に関連のある特徴量を生成して抽出するシステムであり、特徴量選択装置 1 0 0 と、記憶装置 1 1 0 と、プリンタ 1 2 0 とを含んで構成される。特徴量選択装置 1 0 0 と、記憶装置 1 1 0 と、プリンタ 1 2 0 とは、ネットワーク 1 3 0 を介して通信可能に接続されている。

10

【 0 0 1 9 】

特徴量選択装置 1 0 0 は、目的変数に関連のある特徴量を生成する処理などを行う特徴量生成装置の一例であり、ノートパソコン、サーバ装置などである。特徴量選択装置 1 0 0 は、CPU（Central Processing Unit）1 0 1、ROM（Read Only Memory）1 0 2、RAM（Random Access Memory）1 0 3、出力装置 1 0 4、入力装置 1 0 5、通信装置 1 0 6 などを含んで構成される。CPU 1 0 1、ROM 1 0 2、RAM 1 0 3、出力装置 1 0 4、入力装置 1 0 5、および通信装置 1 0 6 は、内部バス 1 0 7 を介して接続されている。

20

【 0 0 2 0 】

特徴量選択装置 1 0 0 の機能（後述の特徴量生成部 5 1 0、相関係数計算部 5 2 0、特徴量抽出部 5 3 0 など）は、例えば、CPU 1 0 1 が ROM 1 0 2 に格納されたプログラムを RAM 1 0 3 に読み出して実行すること（ソフトウェア）により実現されてもよいし、専用の回路などのハードウェアにより実現されてもよいし、ソフトウェアとハードウェアとが組み合わされて実現されてもよい。また、特徴量選択装置 1 0 0 の機能の一部は、特徴量選択装置 1 0 0 と通信可能な他のコンピュータにより実現されてもよい。

【 0 0 2 1 】

出力装置 1 0 4 は、例えばディスプレイであり、各種の情報（例えば、特徴量選択装置 1 0 0 で選択された特徴量）を出力する。入力装置 1 0 5 は、キーボード、ポインティングデバイスなどであり、各種の情報（例えば、後述のユーザ設定 5 4 0 ）を入力する。通信装置 1 0 6 は、例えば、NIC（Network Interface Card）から構成され、記憶装置 1 1 0、プリンタ 1 2 0 との通信時における制御を行う。

30

【 0 0 2 2 】

記憶装置 1 1 0 は、例えば、ストレージ装置であり、各種のデータを格納する DB（data base）1 1 1 を備える。DB 1 1 1 には、例えば、本実施の形態に係るデータを格納する入力テーブルが格納される。なお、入力テーブルについては、図 2 ～ 図 4 を用いて説明する。

【 0 0 2 3 】

プリンタ 1 2 0 は、各種の情報（例えば、特徴量選択装置 1 0 0 で選択された特徴量）を紙などの媒体に印刷する。

40

【 0 0 2 4 】

なお、図 1 では、特徴量選択装置 1 0 0 で選択された特徴量を出力する出力先の例として、出力装置 1 0 4 およびプリンタ 1 2 0 を示したが、これらに限られるものではない。例えば、出力先としては、特徴量選択装置 1 0 0 と通信可能な他のコンピュータであってもよいし、特徴量を利用するソフトウェアであってもよい。

【 0 0 2 5 】

図 2 ～ 図 4 は、特徴量抽出を説明するためのデータの一例（入力テーブル、特徴量テーブル、相関テーブル、特徴量抽出結果など）を示す図である。なお、以下では、入力テーブルは、記憶装置 1 1 0 の DB 1 1 1 に記憶され、特徴量テーブル、相関テーブル、および特徴量抽出結果は、特徴量選択装置 1 0 0 の RAM 1 0 3 に記憶される場合を例に挙げて

50

説明するが、これに限られるものではない。本実施の形態に係るデータは、任意の記憶装置に記憶可能である。

【 0 0 2 6 】

図 2 に示すように、入力テーブル 2 1 0 は、目的変数のカラム（例えば、売上高 2 1 1）と、説明変数のカラム（例えば、イベント A 2 1 2 ~ イベント N 2 1 4）とを含んで構成される。

【 0 0 2 7 】

目的変数の値は、任意の実数である。説明変数の値は、ある事象が起きたか起きなかったかを示す「 0 」または「 1 」の論理値である。

【 0 0 2 8 】

相関テーブル 2 2 0 は、入力テーブル 2 1 0 の全ての説明変数に対して、目的変数との相関（寄与度）を計算した値（相関係数）を格納するテーブルである。相関係数は、基本的には、単位が無く、「 - 1 」から「 1 」の間の実数値をとり、相関係数が正のときは 2 つの確率変数には、正の相関があることを示し、相関係数が負のときは 2 つの確率変数には、負の相関があることを示す。

【 0 0 2 9 】

相関テーブル 2 2 0 に格納された相関係数 R のうち、所定のしきい値（例えば、後述の相関係数目標値 TH_R ）以上の相関係数 R を持つ特徴量が特徴量抽出結果 2 3 0 として出力される。

【 0 0 3 0 】

図 2 に示す例では、イベント B 2 1 3 が特徴量抽出結果 2 3 0 として出力され、これは、イベント B 2 1 3 が起きた場合に、目的変数である「売上高」が増えることを示す分析結果となる。

【 0 0 3 1 】

また、図 3 に示すように、説明変数群に対して、複数の説明変数の論理積をとった値を新たな特徴量として生成し、生成した特徴量に対して相関分析を実施することもある。

【 0 0 3 2 】

図 2 と図 3 との処理の違いは、主に、入力テーブル 3 1 0 から特徴量テーブル 3 2 0 が生成されることである。なお、その後、特徴量テーブル 3 2 0 に基づいて相関テーブル 3 3 0 が生成され、特徴量抽出結果 3 4 0 が出力される。

【 0 0 3 3 】

特徴量テーブル 3 2 0 には、入力テーブル 3 1 0 の説明変数群に対して、1 つ以上の説明変数の論理積による組合せを求めた特徴量が含まれる。例えば、入力テーブル 3 1 0 の説明変数であるイベント A 3 1 1 とイベント B 3 1 2 との論理積による組合せをとった値は、特徴量テーブル 3 2 0 のイベント { A , B } 3 2 1 になり、イベント A 3 1 1 とイベント B 3 1 2 とが同時に起こったことを示す特徴量 { A , B } になっている。

【 0 0 3 4 】

さらに図 3 に示す特徴量にあるイベントが発生したか発生しなかったかを示す「 0 」または「 1 」の論理値ではなく、あるイベントが発生する確率（発生確率）を示す値に置き換え、一般化して考えることもできる。

【 0 0 3 5 】

説明変数および特徴量については、図 3 に示す入力テーブル 3 1 0 と特徴量テーブル 3 2 0 とでは論理値のみしか扱っていなかったが、図 4 に示す入力テーブル 4 1 0 と特徴量テーブル 4 2 0 とでは、発生確率（「 0 」 ~ 「 1 」）を扱う。なお、確率であるため、特徴量同士の論理積による組合せは、それぞれの確率値の積で計算できる。

【 0 0 3 6 】

特徴量選択システム 1 には、図 2 ~ 図 4 の何れの入力テーブル 2 1 0 , 3 1 0 , 4 1 0 にも適用可能であるが、以下では、図 4 に示す入力テーブル 4 1 0 を用いた場合を例に挙げて、特徴量選択システム 1 に係る処理を説明する。

【 0 0 3 7 】

10

20

30

40

50

図 5 は、特徴量選択システム 1 の全体フローを説明するための図である。図 5 に示すように、特徴量選択システム 1 における入力データは、入力テーブル 4 1 0 およびユーザ設定 5 4 0 である。また、出力データは、特徴量抽出結果 4 4 0 である。さらに、中間データとして特徴量テーブル 4 2 0 と相関テーブル 4 3 0 とがある。なお、入力テーブル 4 1 0、特徴量テーブル 4 2 0、相関テーブル 4 3 0、および特徴量抽出結果 4 4 0 については、上述した入力テーブル 3 1 0、特徴量テーブル 3 2 0、相関テーブル 3 3 0、および特徴量抽出結果 3 4 0 と同様である。

【 0 0 3 8 】

ユーザ設定 5 4 0 には、最大組合せ数 N と、相関係数目標値 TH_R とが含まれる。最大組合せ数 N は、最大いくつの説明変数の組合せからなる特徴量を計算するかを示す。相関係数目標値 TH_R は、計算された特徴量のうち目的変数に対する相関係数がいくつ以上の特徴量を抽出するかを示す。最大組合せ数 N および相関係数目標値 TH_R は、分析を始める前にユーザが事前設定する値である。

10

【 0 0 3 9 】

特徴量生成部 5 1 0 は、入力テーブル 4 1 0 に含まれる説明変数から、特徴量テーブル 4 2 0 を生成する。例えば、特徴量生成部 5 1 0 は、目的変数に係る複数の説明変数において組合せを取って特徴量を生成する際、目的変数との相関係数がしきい値（例えば、相関係数目標値 TH_R ）以上になる特徴量を生成する。

【 0 0 4 0 】

相関係数計算部 5 2 0 は、特徴量テーブル 4 2 0 に含まれる全ての特徴量に対して、目的変数との相関係数を計算し、相関テーブル 4 3 0 に出力する。

20

【 0 0 4 1 】

特徴量抽出部 5 3 0 は、相関テーブル 4 3 0 に含まれる特徴量のうち、相関係数がユーザ設定 5 4 0 に含まれる相関係数目標値 TH_R 以上の特徴量を抽出し、特徴量抽出結果 4 4 0 とする。

【 0 0 4 2 】

従来の方では、特徴量生成部 5 1 0 において、最大組合せ数 N 以下の説明変数の特徴量を全て網羅的に計算していたため、組合せ爆発により計算量が膨大になるといった問題がある。

【 0 0 4 3 】

本実施の形態における特徴量選択システム 1 は、かかる問題を解決するための構成として、特徴量生成部 5 1 0 を採用する。特徴量生成部 5 1 0 が実行する処理について図 6 および図 7 を用いて説明する。

30

【 0 0 4 4 】

図 6 は、特徴量生成部 5 1 0 の処理の流れを説明するための模式図である。図 7 は、特徴量生成部 5 1 0 の特徴量生成処理に係るフローチャートの一例を示す図である。

【 0 0 4 5 】

図 6 に示すように、特徴量生成部 5 1 0 は、一度に全ての組合せを網羅計算せずに、1 組合せ、2 組合せ、3 組合せ、・・・、と段階的に組合せ数を増やしながら、特徴量を生成（計算）することを特徴とする。

40

【 0 0 4 6 】

まず、特徴量生成部 5 1 0 は、入力テーブル 4 1 0 に含まれる全ての説明変数を 1 組合せリスト F_1 （1 組合せリスト 6 1 0）とする。図 6 の例では、イベント A ~ イベント E の発生確率を示す特徴量が該当する。

【 0 0 4 7 】

次に、特徴量生成部 5 1 0 は、1 組合せリスト 6 1 0 に含まれる全ての特徴量に対して、他の任意の特徴量と組合せを取った際に、目的変数との間で取り得る相関係数が、ユーザ設定 5 4 0 に含まれる相関係数目標値 TH_R 以上になるか否かを判定する。また、特徴量生成部 5 1 0 は、判定結果が相関係数目標値 TH_R 未満のものを除去し、除去の結果残った 1 組合せリスト 6 1 0 内の特徴量を、1 組合せ候補リスト C_1 （1 組合せ候補リ

50

スト620)に格納する。なお、図6に示す例では、特徴量生成部510は、1組合せリスト610から1組合せ候補リスト620を生成する際、イベントBの特徴量{B}、イベントEの特徴量{E}を除去している。

【0048】

次に、特徴量生成部510は、1組合せ候補リスト620に含まれる特徴量同士の全ての組合せを作り、2組合せリストF__2(2組合せリスト630)に格納する。1組合せ候補リスト620から2組合せリスト630を生成する際、1組合せ候補リスト620の3つの特徴量{A}、{C}、{D}から2つを取り出した組合せである2組合せリスト630として特徴量{A,C}、{A,D}、{C,D}を作っている。例えば、特徴量{A,C}は、イベントAとイベントCとが同時に起こったことを示している。

10

【0049】

次に、特徴量生成部510は、1組合せ候補リスト620から2組合せリスト630を生成する場合と同様に、2組合せリスト630に含まれる全ての特徴量に対して、他の任意の特徴量と組合せを取った際に、目的変数との間で取り得る相関係数が、相関係数目標値TH__R以上になるか否かを判定する。特徴量生成部510は、判定結果が相関係数目標値TH__R未満の特徴量を除去し、除去の結果残った2組合せリスト630内の特徴量を、2組合せ候補リストC__2(2組合せ候補リスト640)に格納する。なお、図6に示す例では、特徴量生成部510は、2組合せリスト630から2組合せ候補リスト640を生成する際、特徴量{A,C}を除去している。

【0050】

20

次に、特徴量生成部510は、2組合せ候補リスト640と1組合せ候補リスト620とに含まれる特徴量の全て組合せを計算し、計算の結果できた3組合せの特徴量を、3組合せリストF__3(3組合せリスト650)に格納する。特徴量生成部510は、3組合せリスト650を生成する際、2組合せ候補リスト640の特徴量{A,D}、{C,D}と1組合せ候補リスト620の特徴量{A}、{C}、{D}とから、それぞれ1つずつ特徴量を取り出し、論理積の組合せをとった、3組合せリスト650の特徴量{A,C,D}を生成している。

【0051】

最後に、特徴量生成部510は、1組合せリスト610、2組合せリスト630、3組合せリスト650の和集合を取った全組合せリストF__ALL(全組合せリスト660)の特徴量{A}、{B}、{C}、{D}、{E}、{A,C}、{A,D}、{C,D}、{A,C,D}を特徴量生成結果として出力する。

30

【0052】

このように、特徴量生成部510が段階的に組合せを生成し、次の組合せの計算に進む際、特徴量候補を判定することで枝刈りする構成について述べたが、本実施の形態は、これに限られるものではない。例えば、特徴量生成部510は、全ての段階で枝刈りをすることなく、所定の段階から進む際に枝刈り(例えば、1組合せリストF__1のみから不要なものを枝刈り、2組合せリストF__2のみから不要なものを枝刈り、1組合せリストF__2と3組合せリストF__3から不要なものを枝刈り等)するようにしてもよい。

【0053】

40

つまり、特徴量生成部(例えば、特徴量生成部510)は、上記複数の説明変数から生成し得る複数の特徴量(例えば、1組合せリスト610の複数の特徴量、2組合せリスト630の複数の特徴量など)の各々について、一の特徴量(例えば、特徴量{A})と上記一の特徴量とは異なる任意の他の特徴量(例えば、特徴量{*})との組合せを取ったときの特徴量(例えば、特徴量{A,*})と上記目的変数との間で取り得る相関係数が上記しきい値(例えば、相関係数目標値TH__R)以上になるか否かを判定し、上記複数の特徴量から上記しきい値未満の特徴量を除去し、除去して残った特徴量同士の組合せ(例えば、1組合せ候補リスト620と1組合せ候補リスト620との全組合せ、2組合せ候補リスト640と1組合せ候補リスト620との全組合せなど)を取った特徴量を新たな特徴量(例えば、2組合せリスト630、3組合せリスト650)として生成することを

50

特徴とする。

【 0 0 5 4 】

上記構成では、一の特徴量の状態を見て、他の特徴量との組合せの相関係数がしきい値を超えるかを事前に判定し、複数の特徴量からしきい値未満の特徴量を除去し、除去して残った特徴量同士の組合せを取った特徴量を新たな特徴量として生成するので、例えば、目的変数に対する相関の大きい特徴量を生成するための計算量を削減することができる。

【 0 0 5 5 】

図 7 を用いて特徴量生成部 5 1 0 が実行する処理について説明する。本フローチャートにおいて、 k は、現在着目している組合せ数を示すものである。

【 0 0 5 6 】

まず、特徴量生成部 5 1 0 は、 $k = 1$ として k を初期化する（ステップ S 7 0 1 ）。

【 0 0 5 7 】

ステップ S 7 0 2 では、特徴量生成部 5 1 0 は、入力テーブル 4 1 0 の全ての説明変数を 1 組合せリスト F_1 に登録する。

【 0 0 5 8 】

ステップ S 7 0 3 では、特徴量生成部 5 1 0 は、現在着目している組合せ数 k が、ユーザ設定 5 4 0 の最大組合せ数 N に達しているか否かを判定する。特徴量生成部 5 1 0 は、達していないと判定した場合、ステップ S 7 0 4 に処理を移し、達していると判定した場合、ステップ S 7 0 8 に処理を移す。

【 0 0 5 9 】

ステップ S 7 0 4 では、特徴量生成部 5 1 0 は、 k 組合せリスト F_k に含まれる全ての特徴量に対して、相関見込み値 R_max を計算する。ここでは、特徴量生成部 5 1 0 は、実際に特徴量を組合せる前に、組合せ後の相関係数の上限値（例えば、相関見込み値 R_max ）を計算している。なお、相関見込み値 R_max の概念および計算方法については後述する。

【 0 0 6 0 】

ステップ S 7 0 5 では、特徴量生成部 5 1 0 は、計算した相関見込み値 R_max がユーザ設定 5 4 0 の相関係数目標値 TH_R 以上であるか否かを判定する。特徴量生成部 5 1 0 は、相関見込み値 R_max が相関係数目標値 TH_R 以上の特徴量を、 k 組合せリスト F_k から、 k 組合せ候補リスト C_k にコピーする。

【 0 0 6 1 】

ステップ S 7 0 6 では、特徴量生成部 5 1 0 は、1 組合せ候補リスト C_1 に含まれる特徴量と、現在着目している k 組合せ候補リスト C_k に含まれる特徴量との全ての組合せ（リストの直積）を計算し、計算結果を $(k + 1)$ 組合せリスト F_k に登録する。このとき、特徴量生成部 5 1 0 は、組合せ数が $(k + 1)$ の特徴量を $(k + 1)$ 組合せリスト F_k に登録し、組合せ数が $(k + 1)$ 未満の特徴量については、 $(k + 1)$ 組合せリスト F_k に登録しない。

【 0 0 6 2 】

例えば、1 組合せ候補リスト C_1 の特徴量が $\{A\}$ 、 $\{D\}$ 、 $\{E\}$ であり、 k 組合せ候補リスト C_k （ここでは、 $k = 2$ の場合を例に挙げて説明する。）の特徴量が $\{A, B\}$ 、 $\{A, D\}$ 、 $\{D, E\}$ である場合、1 組合せ候補リスト C_1 と 2 組合せ候補リスト C_2 の直積を計算すると、

F_3

$= [\{A, A, B\}, \{A, A, D\}, \{A, D, E\}, \{D, A, B\}, \{D, A, D\}, \{D, D, E\}, \{E, A, B\}, \{E, A, D\}, \{E, D, E\}]$

$= [\{A, B\}, \{A, D\}, \{A, D, E\}, \{A, D, B\}, \{A, D\}, \{D, E\}, \{A, B, E\}, \{A, D, E\}, \{D, E\}]$

$= [\{A, B\}, \{A, D\}, \{D, E\}, \{A, B, E\}, \{A, D, E\}]$

となる。組合せ数が「3」未満の特徴量は、登録しないため（1 組合せ数、2 組合せ数については登録しているため）、 $F_3 = [\{A, B, E\}, \{A, D, E\}]$ となる。

10

20

30

40

50

【 0 0 6 3 】

ステップ S 7 0 7 では、特徴量生成部 5 1 0 は、現在着目している組合せ数 k を $k + 1$ として、1 つインクリメントし、ステップ S 7 0 3 に処理を移す。なお、ステップ S 7 0 3 では、特徴量生成部 5 1 0 は、組合せ数 $k =$ 最大組合せ数 N となっている場合、ステップ S 7 0 8 に処理を移す。

【 0 0 6 4 】

ステップ S 7 0 8 では、特徴量生成部 5 1 0 は、これまで計算した組合せリスト $F_1 \sim F_N$ の和集合を求め、特徴量生成結果 F_ALL として出力する。

【 0 0 6 5 】

このように、特徴量生成部（例えば、特徴量生成部 5 1 0）は、説明変数の組合せ数が N （ $N - 2$ の整数）の特徴量を生成する際、説明変数の組合せ数が k （ $1 \leq k \leq N$ の整数）の複数の特徴量の各々について、一の特徴量と上記一の特徴量とは異なる任意の他の特徴量との組合せを取ったときの特徴量と上記目的変数との間で取り得る相関係数（例えば、相関見込み値 R_max ）が上記しきい値以上になるか否かを判定し、上記複数の特徴量から上記しきい値未満の特徴量を除去し、除去して残った説明変数の組合せ数が「1」の特徴量と除去して残った組合せ数が k の特徴量との組合せを取った特徴量を説明変数の組合せ数が $k + 1$ の新たな特徴量として生成する処理を、説明変数の組合せ数が N の特徴量を生成するまで k が 1 から順に繰り返すことを特徴とする。

10

【 0 0 6 6 】

かかる処理によれば、組合せ数を段階的に増やして特徴量を生成する際に、以降の段階でどのように組合せてもしきい値を超えない特徴量が事前に枝刈りされるので、効率よく特徴量を生成することができる。

20

【 0 0 6 7 】

次に、相関見込み値 R_max の計算方法について説明する。図 8 および図 9 を用いて、相関見込み値 R_max の計算ポイントについて説明する。

【 0 0 6 8 】

図 8 は、特徴量の発生確率を横軸 x に、目的変数の値を縦軸 y にとった散布図の一例を示す図である。図 8 に示す散布図は、特徴量テーブル 4 2 0 に示す特徴量ごとに生成され、特徴量の各レコードの値が、散布図上の点群の 1 点 1 点に対応する。散布図における点 (y_i, x_i) は、 i 番目のレコードに対応する点を示す。 y_i は、任意の実数であり、 x_i は、「0.0」～「1.0」の値である。

30

【 0 0 6 9 】

図 8 では、特徴量 $\{A\}$ が存在し、さらに特徴量 $\{A\}$ と他の任意の特徴量と組み合わせを取ることでできる特徴量 $\{A, *\}$ があることを考える。

【 0 0 7 0 】

グラフ 8 1 0 は、組合せ前の特徴量 $\{A\}$ における点群 (y_i, x_i) を示す。グラフ 8 2 0 は、組合せ後の特徴量 $\{A, *\}$ における点群 (y_i, x_i^*) を示す。本特徴量選択システム 1 では、特徴量同士の論理積を取ることにによる組合せ生成を考えているため、 A かつ B が同時に起こる確率は、 A または B がそれぞれ起こる確率よりも小さくなる。言い換えれば、ある特徴量 $\{A\}$ と特徴量 $\{B\}$ とを組合せた特徴量 $\{A, B\}$ の発生確率は、特徴量 $\{A\}$ の単独の発生確率よりも小さな値となり、かつ、特徴量 $\{B\}$ の単独の発生確率よりも小さな値となる。

40

【 0 0 7 1 】

図 8 に示す散布図においては、組合せ後のグラフ 8 2 0 の点群 (y_i, x_i^*) は、組合せ前のグラフ 8 1 0 の点群 (y_i, x_i) よりも、 x 軸の小さな方向にしか存在せず、 y 軸方向には変化がない。すなわち、 $0 \leq x_i^* \leq x_i$ 、と表すことができる。この性質を用いて、組合せ前の特徴量 $\{A\}$ に対する点群 (y_i, x_i) から、他の任意の特徴量との組合せをとることで作った特徴量 $\{A, *\}$ に対する点群 (y_i, x_i^*) における、目的変数 y_i と特徴量の発生確率 x_i^* との相関係数 $R(y_i, x_i^*)$ を考え、この相関係数 $R(y_i, x_i^*)$ が取り得る大きな値を見積もることを考える。

50

【 0 0 7 2 】

図 9 は、図 8 同様に、特徴量の発生確率を横軸 x に、目的変数の値を縦軸 y にとった散布図の一例を示す図である。

【 0 0 7 3 】

ここで、グラフ 9 1 0 のような組合せ前のある特徴量 $\{A\}$ に対する点群 (y_i, x_i) があり、点群 (y_i, x_i) の発生確率 x_i を $0 \leq x_i^* \leq x_i$ の条件で移動させた際に、相関係数 $R(y_i, x_i^*)$ になるべく大きくなる移動パターンを考えると、例えば、グラフ 9 2 0 に示すような、ある移動境界直線 $y = ax + b$ への移動が考えられる。

【 0 0 7 4 】

より具体的には、移動前の点群 (y_i, x_i) を、移動境界直線 $y = ax + b$ を境として、グラフ上の右側 (x 軸の大きい側) にある場合には、移動境界直線上に移動し、グラフ 9 2 0 上の左側 (x 軸の小さい側) にある場合には、移動しない。また、 $x < 0$ の範囲には移動できず、既に $x_i = 0$ の点は、動かないことになる。数式で表せば、以下のようになる。

(数式 1)

```
if  $y_i > b$  and  $x_i \leq (y_i - b) \div a$  then  $x_i^* = x_i$ 
if  $y_i > b$  and  $x_i > (y_i - b) \div a$  then  $x_i^* = (y_i - b) \div a$ 
if  $y_i \leq b$  then  $x_i^* = 0$ 
```

【 0 0 7 5 】

すなわち、本特徴量選択システム 1 では、例えば (数式 1) を用いて、移動前の点群 (y_i, x_i) から、移動後の点群 (y_i, x_i^*) が取り得る相関係数のなるべく大きな値を見積もることで、図 7 のステップ S 7 0 4、ステップ S 7 0 5 に示すように、組合せ前のある特徴量 $\{A\}$ を、ある任意の特徴量と組み合わせた際の特徴量 $\{A, *\}$ と、目的変数との相関係数である相関見込み値 R_max が、目標値 TH_R を超え得るか否かを事前に判断することができる。

【 0 0 7 6 】

また、以上の説明は、ある特徴量における目的変数と発生確率との相関係数にて、正の相関係数の大きい値を見積もる方法であるが、図 1 0 のように、正相関の場合と同様の手順で (数式 1) の代わりに (数式 2) を用いて、負の相関係数の大きい値を見積もることができる。

(数式 2)

```
if  $y_i < b$  and  $x_i \leq (y_i - b) \div a$  then  $x_i^* = x_i$ 
if  $y_i < b$  and  $x_i > (y_i - b) \div a$  then  $x_i^* = (y_i - b) \div a$ 
if  $y_i \geq b$  then  $x_i^* = 0$ 
```

【 0 0 7 7 】

なお、グラフ 1 0 1 0 は、組合せ前の特徴量 $\{A\}$ における点群 (y_i, x_i) を示すグラフである。グラフ 1 0 2 0 は、組合せ後の特徴量 $\{A, *\}$ における点群 (y_i, x_i^*) を示すグラフである。

【 0 0 7 8 】

図 1 1 は、相関がなるべく大きくなるときの移動先の点群 (y_i, x_i^*) の考え方について補足するための図である。

【 0 0 7 9 】

ここで、グラフ 1 1 1 0 のように点群 (y_i, x_i) が散布しており、 y_i の平均 $mean(y_i)$ と x_i の平均 $mean(x_i)$ にて分割される 4 つの象限 I ~ IV がある状態を考える。また、 x_i の平均 $mean(x_i)$ は、ある程度小さな状態を想定する。

【 0 0 8 0 】

図 1 1 に示す一般的な相関係数の計算式 1 1 2 0 に基づいて、グラフ 1 1 1 0 の状態から正の相関がなるべく大きくなる状態を考えると、象限 I および象限 III に点がなるべく多く存在し、象限 II および象限 IV に点がなるべく少なく存在する状態となる。さらに、象限 I および象限 III の点は、なるべく傾きの大きい直線に近い状態に並んでいる状

10

20

30

40

50

態となる。一方、前述のように移動前の点群 (y_i, x_i) と移動後の点群 (y_i, x_i^*) には、 $0 < x_i^* < x_i$ の関係があり、これらを整理すると、各象限 I ~ IV において、移動後の点群 (y_i, x_i^*) の相関がなるべく大きくなる状態は、例えば、下記のように整理できる（グラフ 1130 に図示）。

【0081】

象限 I : 各点がなるべく傾きの大きい直線に近い状態に並ぶ。

象限 II : 点が存在しない（すべて象限 III に移動）。

象限 III : 各点がなるべく傾きの大きい直線に近い状態に並ぶ。ただし、 $x_i^* < 0$ にはならないため、 $x_i^* = 0$ で下げ止まる。

象限 IV : 点がなるべく存在しない。ただし、移動前に $x_i = 0$ に存在する点が移動できないのでそのまま存在する。

10

【0082】

上記の性質より、正相関がなるべく大きくなるための点群 (y_i, x_i) から (y_i, x_i^*) の移動パターンとしては、例えば、点 $(y_i$ の平均, x_i の平均) を通り、かつ、傾きのなるべく大きい移動境界直線 $y = ax + b$ を考え、この移動境界直線に沿うように点群が移動した状態と考えることができる。

【0083】

ただし、移動パターンは、点 $(y_i$ の平均, x_i の平均) を通る移動境界直線に限られるものではない。例えば、図 14 に示すように移動境界直線を求めてもよいし、他の方法により移動境界直線を求めてもよい。また、移動パターンは、直線に限られるのではなく、曲線であってもよい。

20

【0084】

また、上述の実施の形態においては、相関がなるべく大きくなるように点群 (y_i, x_i) を移動するために移動境界直線を算出する場合について述べたが、本実施の形態は、これに限られるものではない。例えば、特徴量生成部 510 は、移動境界直線を算出することなく点群 (y_i, x_i) を移動してもよい。上記の性質によると、より広義には、特徴量生成部 510 は、相関がなるべく大きくなるように、点群 (y_i, x_i) において、 y_i は、変更することなく、所定の値 m （例えば、 $0 < m < \text{平均mean}(x_i)$ ）より大きい x_i を所定の値 m に移動（例えば、平均 $\text{mean}(x_i)$ より大きい x_i を平均 $\text{mean}(x_i)$ に移動）するようにしてもよい。

30

【0085】

つまり、特徴量生成部（例えば、特徴量生成部 510）は、上記複数の特徴量から上記しきい値未満の特徴量を除去する際、上記目的変数の値と上記一の特徴量の値とを示す点群に基づいて、上記一の特徴量と上記他の特徴量との組合せを取ったときの特徴量と上記目的変数との間で取り得る相関係数が十分に大きくなるように、上記目的変数の値を変更することなく上記一の特徴量の値を変更することで上記点群を移動し、移動した点群と上記目的変数との相関係数である相関見込み値を計算し、計算した相関見込み値が上記しきい値以上であるか否かを判定することを特徴とする。

【0086】

かかる構成によれば、一の特徴量と他の特徴量との組合せを取ったときの特徴量と目的変数との間で取り得る相関係数を十分に大きくすることができる。

40

【0087】

図 12 は、相関見込み値 R_{max} の計算処理に係るフローチャートの一例を示す図である。

【0088】

ステップ S1201 では、特徴量生成部 510 は、相関見込み値 R_{max} の計算対象の特徴量における目的変数と発生確率との点群 $D = (y_i, x_i)$ を生成する。

【0089】

ステップ S1202 では、特徴量生成部 510 は、点群 D に基づいて、移動境界直線 $y = ax + b$ を計算する。移動境界直線のパラメータ a および b の定め方の例については、図

50

13 および図14を用いて後述する。

【0090】

ステップS1203では、特徴量生成部510は、現在の点群 $D = (y_i, x_i)$ から、移動先の点群 $D^* = (y_i, x_i^*)$ を(数式1)または(数式2)を用いて計算する。

【0091】

ステップS1204では、特徴量生成部510は、移動先の点群 D^* における目的変数 y_i と発生確率 x_i^* との相関係数を、計算対象の特徴量の相関見込み値 R_max とし、計算処理を終了する。なお、特徴量生成部510と相関係数計算部520とは、同じ計算式(例えば、図11に示す計算式1120)を用いて相関係数を算出する。

【0092】

次に、移動境界直線 $y = ax + b$ のパラメータである傾き a および切片 b の計算例について、図13および図14を用いて説明する。

【0093】

図13は、移動境界直線の計算処理に係るフローチャートの一例を示す図である。

【0094】

ステップS1301では、特徴量生成部510は、点 α を求める。特徴量生成部510は、正相関の相関見込み値 R_max を計算する場合、 y_i の最大値を用いて $\alpha = (max(y_i), \Delta x)$ とし、負相関の相関見込み値 R_max を計算する場合、 y_i の最小値を用いて $\alpha = (min(y_i), \Delta x)$ とする。 Δx は、非常に小さな値とし、例えば、 $\Delta x = 0.0001$ などを与える。グラフ1110で説明したように、点 α は、移動境界直線の傾きがある程度十分に大きくなるように配置する点である。

【0095】

図14は、正相関の相関見込み値 R_max を計算する場合の例を示す。グラフ1410では、点 α を図示している。

【0096】

ステップS1302では、特徴量生成部510は、点 β の候補群 $\{\beta_n\}$ を計算する。点 β は、移動境界直線の切片 b となる値である。特徴量生成部510は、正相関の相関見込み値 R_max を計算する場合、 y_i の平均 $mean(y_i)$ から y_i の最小値 $min(y_i)$ の間を n 刻みに当分割した点とする。負相関の相関見込み値 R_max を計算する場合、 y_i の最大値 $max(y_i)$ から y_i の平均 $mean(y_i)$ の間を n 刻みに当分割した点とする。

【0097】

グラフ1420では、 $mean(y_i)$ から $min(y_i)$ の間を5分割した $\{\beta_1 \sim \beta_5\}$ が点 β の候補として図示している。

【0098】

ステップS1303では、特徴量生成部510は、2点(点 α と点 β の候補群 $\{\beta_n\}$ の各々)を通る直線を求め、移動境界直線候補 $\{L_n\}$ とする。

【0099】

ステップS1304では、特徴量生成部510は、点群 $D = (y_i, x_i)$ を、それぞれの移動境界直線候補 $\{L_n\}$ に沿って移動した、移動先の点群候補 $\{D_n\}$ を計算する。移動境界直線候補 $\{L_n\}$ とそれらに対応した移動先の点群候補 $\{D_n\}$ が計算される。

【0100】

グラフ1430の例では、 L_1 および D_1 について計算した移動境界直線候補 L_1 を図示している。

【0101】

ステップS1305では、特徴量生成部510は、各移動先の点群候補 $\{D_n\}$ の相関係数の絶対値が最大となるときの、対応する移動境界直線候補 L_n を移動境界直線 L と決定する。

【0102】

グラフ1440では、移動境界直線候補 L_1 が移動境界直線 L として決定されていること

10

20

30

40

50

が示されている。

【 0 1 0 3 】

このように、特徴量生成部（例えば、特徴量生成部 5 1 0）は、上記目的変数の値（例えば、売上高の値）と上記一の特徴量の値（例えば、特徴量 { A } の発生確率）とを示す点群（例えば、(y_i , x_i)）に基づいて、上記一の特徴量と上記他の特徴量との組合せを取ったときの特徴量（例えば、特徴量 { A , * }）と上記目的変数との間で取り得る相関係数（例えば、相関見込み値 R_{max}）が十分に大きくなる（例えば、上限となる）ための上記点群の移動パターンを示す移動境界直線を計算し、計算した移動境界直線（例えば、移動境界直線 L）に基づいて上記点群を移動することを特徴とする。

【 0 1 0 4 】

かかる構成によれば、例えば、上限により近い相関係数を算出でき、枝切りの精度を高めることができるので、より適切に特徴量を生成することができる。

【 0 1 0 5 】

[本実施の形態の効果]

本特徴量選択システム 1 によれば、より少ない計算時間で、目的変数に対する相関の高い特徴量を抽出することで、データ分析者の分析サイクルの短縮が可能になる。

【 0 1 0 6 】

また、本特徴量選択システム 1 によれば、特徴量の計算時の組合せ爆発を回避することで、データ分析者の不要な事前作業を軽減できる。

【 0 1 0 7 】

また、本特徴量選択システム 1 によれば、より少ない計算回数および特徴量候補で、目的変数に対する相関の高い特徴量を抽出することで、メモリなどの計算コストを低減できる。

【 0 1 0 8 】

(2) 他の実施の形態

なお上述実施の形態においては、本発明を特徴量選択システム 1 に適用するようにした場合について述べたが、本発明はこれに限らず、この他種々のシステム、方法、プログラム、記憶媒体などに広く適用することができる。

【 0 1 0 9 】

また上述の実施の形態においては、特徴量選択装置 1 0 0 について図 1 に示す構成を例に挙げて述べたが、本発明はこれに限らず、特徴量選択装置 1 0 0 は、例えば、通信装置 1 0 6 などを備えていなくてもよいし、HDD (Hard Disk Drive) などの他の装置を備えていてもよい。

【 0 1 1 0 】

また上述の実施の形態においては、入力テーブルが記憶装置 1 1 0 (DB 1 1 1) に設けられる場合について述べたが、本発明はこれに限らず、入力テーブルが特徴量選択装置 1 0 0 の記憶装置 (RAM 1 0 3 、 HDD 等) に設けられるようにしてもよい。

【 0 1 1 1 】

また上述の実施の形態においては、最大組合せ数 N は、ユーザにより設定される場合について述べたが、本発明はこれに限らず、最大組合せ数 N は、入力テーブルを参照してカラムの数を特徴量選択装置 1 0 0 が自動で設定するようにしてもよい。

【 0 1 1 2 】

また上述の実施の形態においては、説明の便宜上、X X テーブルを用いて各種のデータを説明したが、データ構造は限定されるものではなく、X X 情報などと表現してもよい。

【 0 1 1 3 】

また、上記の説明において各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、SSD (Solid State Drive) 等の記憶装置、または、IC カード、SD カード、DVD 等の記録媒体に置くことができる。

【 0 1 1 4 】

また上述した構成については、本発明の要旨を超えない範囲において、適宜に、変更したり、組み替えたり、組み合わせたり、省略したりしてもよい。

10

20

30

40

50

【 0 1 1 5 】

上述した構成によれば、目的変数に対する相関の大きい特徴量を適切に生成することができる。

【 符号の説明 】

【 0 1 1 6 】

1 特徴量選択システム、 1 0 0 特徴量選択装置。

10

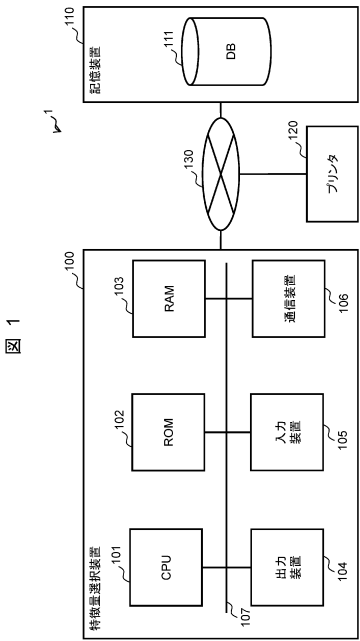
20

30

40

50

【図面】
【図 1】



【図 2】

図 2

入力テーブル

売上高	インストア	インストア	インストア	インストア	インストア
1000	0	1	0	...	インストア
10000	0	1	0	...	0
10000	0	1	0	...	1
12000	0	0	0	...	0
2000	1	0	0	...	0
4000	1	1	1	...	0
600	1	1	1	...	0
5000	0	0	0	...	1
13000	0	0	1	...	0
3000	0	1	0	...	0
30000	0	0	0	...	1
8000	0	0	0	...	1
15000	1	1	1	...	1

目的変数

説明変数

相関テーブル

インストア	インストア	インストア	インストア
0.1	0.4	...	...
...	...	...	...
...	...	...	...

目的変数と各説明変数との相関(寄与度)

特徴量抽出結果

インストア	インストア	インストア	インストア
0.1	0.4	...	...
...	...	...	...
...	...	...	...

R²=TH_R ⇒ {インストア}

【図 3】

図 3

特徴量テーブル

売上高	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア
1000	0	...	...	...	...	...	...	...	...
10000	0	1	0	0	0	0	0	0	0
12000	0	1	0	0	0	0	0	0	0
2000	1	0	0	0	0	0	0	0	0
4000	1	1	0	0	0	0	0	0	0
600	1	1	0	0	0	0	0	0	0
5000	0	0	0	0	0	0	0	0	0
13000	0	0	0	0	0	0	0	0	0
3000	0	1	0	0	0	0	0	0	0
30000	0	0	1	0	0	0	0	0	0
8000	0	0	0	1	0	0	0	0	0
15000	1	1	1	1	1	1	1	1	1

目的変数

説明変数

相関テーブル

インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア
0.1	0.4	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...

目的変数と各説明変数との相関

特徴量抽出結果 R²=TH_R ⇒ {インストア(B), インストア(A,B)}

【図 4】

図 4

特徴量テーブル

売上高	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア
1000	0.8	0.1	...	...	...	...	...	...	...
10000	0.1	1.0	0.2	0.2	...	...	...	...	...
12000	0.2	0.0	0.0	0.7	...	...	...	...	...
2000	0.2	0.0	0.0	0.3	...	...	...	...	...
4000	0.5	0.8	0.0	0.0	...	...	...	...	...
600	1.0	0.7	...	0.0	...	...	...	...	...
5000	0.0	0.1	...	1.0	...	...	...	...	...
13000	0.0	0.2	...	0.9	...	...	...	...	...
3000	0.0	0.6	...	0.0	...	...	...	...	...
30000	0.0	0.1	...	0.8	...	...	...	...	...
8000	0.0	0.0	...	0.7	...	...	...	...	...
15000	0.8	1.0	...	0.1	...	...	...	...	...

目的変数

説明変数

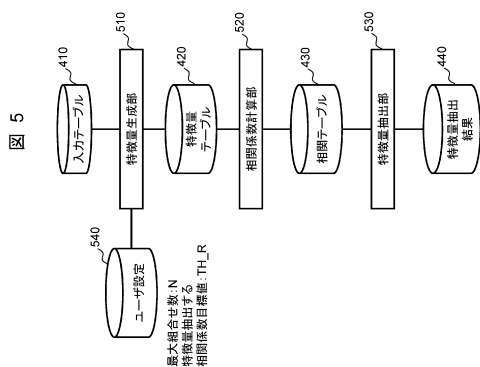
相関テーブル

インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア	インストア
0.1	0.4	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...

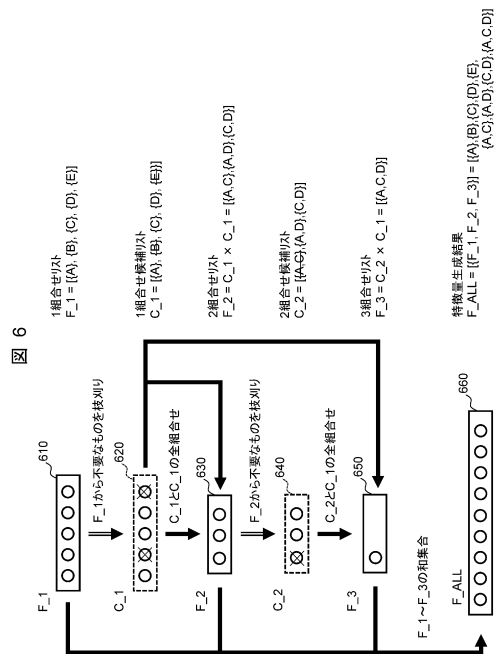
目的変数と各説明変数との相関

特徴量抽出結果 R²=TH_R ⇒ {インストア(B), インストア(A,B)}

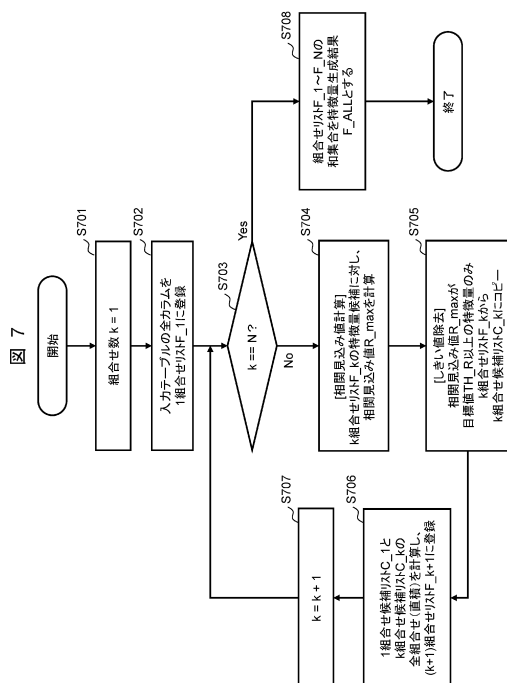
【図 5】



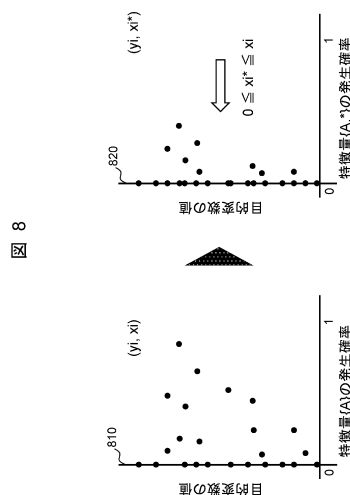
【図 6】



【図 7】



【図 8】



10

20

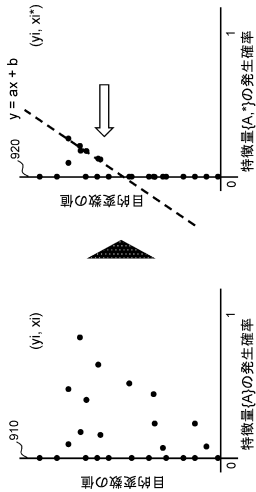
30

40

50

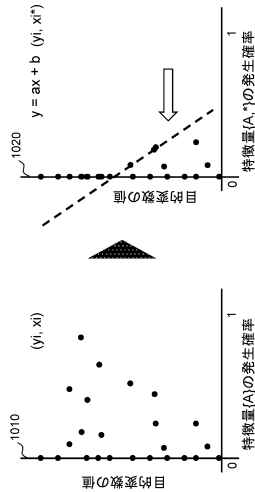
【図 9】

図 9



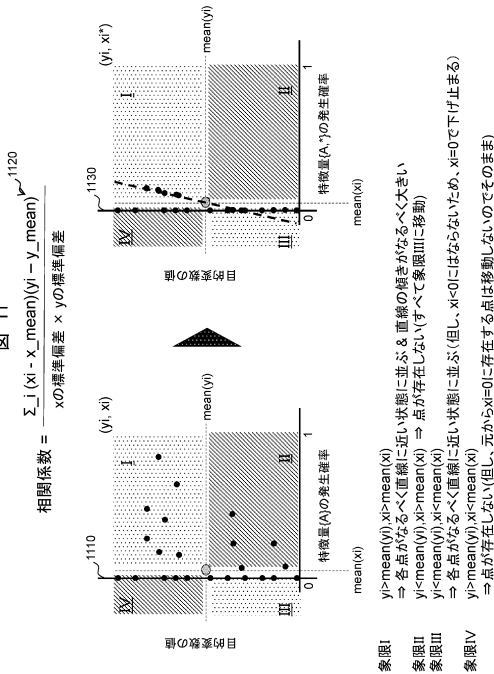
【図 10】

図 10



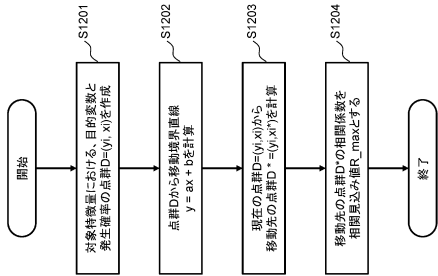
【図 11】

図 11

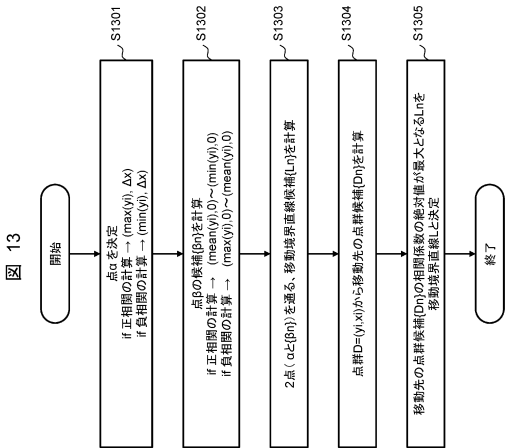


【図 12】

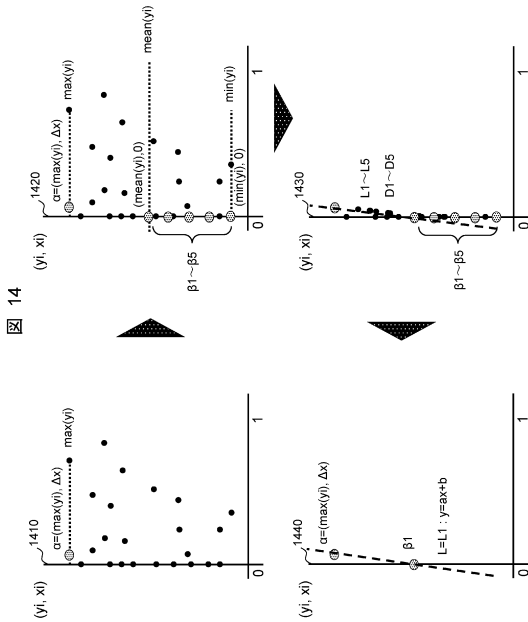
図 12



【図 13】



【図 14】



10

20

30

40

50

フロントページの続き

審査官 白石 圭吾

(56)参考文献 特開平 0 7 - 0 9 3 2 8 4 (J P , A)

特開 2 0 0 6 - 0 3 1 4 4 2 (J P , A)

国際公開第 2 0 1 7 / 2 0 3 6 0 1 (W O , A 1)

(58)調査した分野 (Int.Cl. , D B 名)

G 0 6 F 1 6 / 0 0 - 1 6 / 9 5 8

G 0 6 F 1 7 / 0 0 - 1 7 / 1 8

G 0 6 Q 1 0 / 0 0 - 1 0 / 1 0 ; 3 0 / 0 0 - 3 0 / 0 8 ; 5 0 / 0 0 - 5 0 / 2 0 ;

5 0 / 2 6 - 9 9 / 0 0 ; G 1 6 Z 9 9 / 0 0