

明 細 書

発明の名称：

特徴量生成装置、特徴量生成方法および特徴量生成プログラム

技術分野

[0001] 本発明は、複数のテーブルを結合して特徴量を生成する特徴量生成装置、特徴量生成方法および特徴量生成プログラムに関する。

背景技術

[0002] データマイニングは、大量の情報の中から、これまで未知であった有用な知見を見つける技術である。未知である有用な知見を見つけるためには、より多くの属性の候補を生成することが重要である。具体的には、予測対象である変数（目的変数）に影響を及ぼし得る多くの属性（説明変数）の候補を生成することが重要である。このような多くの候補を生成することにより、予測に役立つ属性がこの候補の中に含まれる可能性を高めることができるからである。

[0003] 例えば、特許文献1には、目的変数を含むターゲットテーブルと、目的変数を含まないソーステーブルとを結合することにより、機械学習処理に用いられる特徴量の候補を生成することが記載されている。特許文献1に記載された方法では、特徴量の候補を生成する処理を、F i l t e r条件、m a p条件およびr e d u c e条件の3つの条件の組合せにより定義することで、特徴量の候補を生成する分析者工数を削減する。

先行技術文献

特許文献

[0004] 特許文献1：国際公開第2017/090475号

発明の概要

発明が解決しようとする課題

[0005] 本発明者は、所定のエリア内における何らかの対象を予測する際、多様な情報源を活用した方が予測精度が向上するという着想を得た。すなわち、複

数の関連する情報源を組み合わせることで情報を得ることが好ましいと考えられる。

[0006] 例えば、特許文献1には、ターゲットテーブルとソーステーブルとの結合条件（すなわち、map条件）に、ターゲットテーブルとソーステーブルに共通に含まれる顧客IDを利用することが例示されている。ここで、本発明者は、分析対象によっては、map条件の候補が複数生成され得ることを見出した。このような分析対象に利用する特徴量の候補を生成しようとした場合、特許文献1に記載された方法を用いても、処理が煩雑になる結果、分析者工数が増大してしまう可能性がある。

[0007] そこで、本発明は、特徴量を生成するための分析者工数を削減できる特徴量生成装置、特徴量生成方法および特徴量生成プログラム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] 本発明による特徴量生成装置は、予測対象および第1の属性を含む第1のテーブルと、第2の属性を含む第2のテーブルとを取得するテーブル取得手段と、第1の属性と第2の属性との類似度の算出に用いられる類似度関数と、その類似度に対する条件とを受け付ける受付手段と、類似度関数を用いて、第1の属性の値と第2の属性の値とにより算出される類似度が、条件を満たす第1の属性の値を含む第1のテーブルのレコードと、第2の属性の値を含む第2のテーブルのレコードとを結合するための結合条件と、第2のテーブルにおける複数のレコードに対する集約方法および集約の対象になる列により表される集約条件とを用いて、第1のテーブルおよび第2のテーブルから、予測対象に影響を及ぼし得る特徴量の候補を生成する特徴量生成手段と、特徴量の候補から、予測に最適な特徴量を選択する特徴量選択手段とを備えたことを特徴とする。

[0009] 本発明による特徴量生成方法は、予測対象および第1の属性を含む第1のテーブルと、第2の属性を含む第2のテーブルとを取得し、第1の属性と第2の属性との類似度の算出に用いられる類似度関数と、その類似度に対する

条件とを受け付け、類似度関数を用いて、第1の属性の値と第2の属性の値とにより算出される類似度が、条件を満たす第1の属性の値を含む第1のテーブルのレコードと、第2の属性の値を含む第2のテーブルのレコードとを結合するための結合条件と、第2のテーブルにおける複数のレコードに対する集約方法および集約の対象になる列により表される集約条件とを用いて、第1のテーブルおよび第2のテーブルから、予測対象に影響を及ぼし得る特徴量の候補を生成し、特徴量の候補から、予測に最適な特徴量を選択することを特徴とする。

[0010] 本発明による結合条件生成プログラムは、コンピュータに、予測対象および第1の属性を含む第1のテーブルと、第2の属性を含む第2のテーブルとを取得するテーブル取得処理、第1の属性と第2の属性との類似度の算出に用いられる類似度関数と、その類似度に対する条件とを受け付ける受付処理、類似度関数を用いて、第1の属性の値と第2の属性の値とにより算出される類似度が、条件を満たす第1の属性の値を含む第1のテーブルのレコードと、第2の属性の値を含む第2のテーブルのレコードとを結合するための結合条件と、第2のテーブルにおける複数のレコードに対する集約方法および集約の対象になる列により表される集約条件とを用いて、第1のテーブルおよび第2のテーブルから、予測対象に影響を及ぼし得る特徴量の候補を生成する特徴量生成処理、および、特徴量の候補から、予測に最適な特徴量を選択する特徴量選択処理を実行させることを特徴とする。

発明の効果

[0011] 本発明によれば、上述した技術的手段により、特徴量を生成するための分析者工数を削減できるという技術的效果を奏する。

図面の簡単な説明

[0012] [図1]本発明による情報処理システムの一実施形態を示すブロック図である。

[図2]コンフィギュレーションファイルの例を示す説明図である。

[図3]データを変換する処理の例を示す説明図である。

[図4]各パラメータと、第1のテーブルおよび第2のテーブルとの関係の例を

示す説明図である。

[図5]距離に基づいてマップパラメータを生成する処理の例を示す説明図である。

[図6]距離に基づいてマップパラメータを生成する他の処理の例を示す説明図である。

[図7]同じエリアに含まれているか否か判断する方法の例を示す説明図である。

[図8]共通エリアか否かに基づいてマップパラメータを生成する処理の例を示す説明図である。

[図9]包含関係に基づいてマップパラメータを生成する処理の例を示す説明図である。

[図10]時間の差異に基づいてマップパラメータを生成する処理の例を示す説明図である。

[図11]テキストの類似性に基づいてマップパラメータを生成する処理の例を示す説明図である。

[図12]構造の類似性に基づいてマップパラメータを生成する処理の例を示す説明図である。

[図13]生成されたマップパラメータの例を示す説明図である。

[図14]距離の統計値を算出するための集約パラメータを生成する処理の例を示す説明図である。

[図15]領域の統計値を算出するための集約パラメータを生成する処理の例を示す説明図である。

[図16]生成された集約パラメータの例を示す説明図である。

[図17]マップパラメータ同士を結合した例を示す説明図である。

[図18]パラメータを組み合わせて特徴量生成関数を生成する方法の例を示す説明図である。

[図19]結合条件を生成する処理の例を示すフローチャートである。

[図20]結合条件を生成する処理の他の例を示すフローチャートである。

[図21]特徴量を生成する処理の例を示すフローチャートである。

[図22]特徴量を生成する処理の他の例を示すフローチャートである。

[図23]本発明による特徴量生成装置の概要を示すブロック図である。

[図24]少なくとも1つの実施形態に係るコンピュータの構成を示す概略ブロック図である。

発明を実施するための形態

[0013] 以下、本発明の実施形態を図面を参照して説明する。

[0014] 本実施形態の情報処理システムは、予測対象（例えば、目的変数）の変数を含む表（以下、第1のテーブルと記すこともある。）と、第1のテーブルと異なる表（以下、第2のテーブルと記すこともある。）とを取得する。以下の説明では、第1のテーブルのことをターゲットテーブルと記すこともあり、第2のテーブルをソーステーブルと記すこともある。また、第1のテーブルおよび第2のテーブルは、それぞれデータの集合を含んでいてもよい。

[0015] 本実施形態では、第1のテーブルおよび第2のテーブルは、観点が共通する属性をそれぞれ含む。観点が共通するとは、その属性のデータの意味的な内容が共通することを表す。なお、データの表現方法は共通であってもよく、異なってもよい。以下、第1のテーブルに含まれる属性を第1の属性と記し、第2のテーブルに含まれる属性を、第2の属性と記す。

[0016] 例えば、観点が共通する属性として、地理的な観点や時間的な観点などが挙げられる。例えば、地理的な観点を属性の値は、以下の4種類の地理的データ型に分類できる。なお、見出しのコロン以下の記載は、データについての構文を表す。

(1) 点P (Point) : $p = (x, y) \in P$

点Pは、(経度, 緯度)の座標として表される。

(2) 多角形G (Polygon) : $g = (b_1, b_2, \dots, b_n) \in G$

多角形Gは、1つの外部境界 b_1 と、0以上の内部境界 $(b_2, \dots, b_n)$ で定義される。ここで、 $b_1 = (p_1, p_2, \dots, p_n)$ (ただし、 $p_1, p_2, \dots, p_n \in P$) は、3点以上の順序として定義される閉じた環の境界である。

(3) 複数多角形 M (Multi Polygon) : $m = (g_1, g_2, \dots, g_n) \in M, g_1, g_2, \dots, g_n \in G$

複数多角形 M は、1 以上の多角形で構成される。

(4) 文字列 S (String) : $s \in S$

文字列で表される住所である。

[0017] また、データ分析に関連する意味的な情報として、データ型と対応付けて分析データ型が定義されてもよい。例えば、上述する地理的観点の場合、多角形 G および複数多角形 M を、領域 (Area) に関する分析データ型と定義し、点 P を、点 (Point) に関する分析データ型と定義してもよい。また、住所に関する文字列を、例えば、国、都市、町、ランドマーク、通りまたはポイントに関する分析データ型と定義してもよい。以下、地理的情報を表す分析データ型のことを、地理的データ型と記すこともある。

[0018] また、例えば、時間的な観点の属性の型 (時間的データ型) は、タイムスタンプ (TimeStamp) 型として定義することができる。

[0019] 以下、観点が共通する属性が地理的な属性の場合、第1のテーブルに含まれる属性を第1の地理的属性、第2のテーブルに含まれる属性を第2の地理的属性と記す。同様に、観点が共通する属性が時間的な属性の場合、第1のテーブルに含まれる属性を第1の時間的属性、第2のテーブルに含まれる属性を第2の時間的属性と記す。他の属性についても同様に記載するものとする。なお、第1の地理的属性は、第1のテーブルのプライマリキーであってもよい。

[0020] なお、上記では、共通する属性が地理的な観点と、時間的な観点を例を示したが、共通する属性は、地理的な観点および時間的な観点到限定されない。共通する属性の例として、他にも、文字列の観点や、構造的な観点などが挙げられる。文字列の観点の属性の値は、例えば、住所などである。また、構造的な観点の属性の値は、例えば、URL (Uniform Resource Locator) や、木構造パスなどである。以下、説明を容易にするため、観点が共通する属性として、主に地理的属性と時間的属性を中心に説明する。

- [0021] 図1は、本発明による情報処理システムの一実施形態を示すブロック図である。本実施形態の情報処理システム100は、入力部10と、ジオコーダ(Geo-Coder)20と、マップパラメータ生成器(Map Parameter Generator)30と、フィルタパラメータ生成器(Filter Parameter Generator)50と、集約パラメータ生成器(Reduce Parameter Generator)60と、記憶部80と、特徴量生成関数生成器(Feature Descriptor Generator)81と、特徴量生成器(Feature Generator)82と、特徴量選択器(Feature Selector)83と、出力部90と、学習部91と、予測部92とを備えている。
- [0022] 入力部10は、第1のテーブルおよび第2のテーブルを取得する。なお、入力部10は、各テーブルを取得することから、入力部10のことをテーブル取得手段とすることができる。入力部10は、第2のテーブルを複数取得してもよい。例えば、記憶部80が第1のテーブルおよび第2のテーブルを記憶している場合、入力部10が記憶部80から第1のテーブルおよび第2のテーブルを取得してもよい。また、入力部10は、通信ネットワーク(図示せず)を介して他のシステムや記憶部から第1のテーブルおよび第2のテーブルを取得してもよい。
- [0023] 例えば、地理的な観点が共通する場合、入力部10は、予測対象および第1の地理的属性を含む第1のテーブルと、第2の地理的属性を含む第2のテーブルとを取得してもよい。また、例えば、時間的な観点が共通する場合、入力部10は、予測対象および第1の時間的属性を含む第1のテーブルと、第2の時間的属性を含む第2のテーブルとを取得してもよい。他にも、入力部10は、予測対象および第1の文字列属性を含む第1のテーブルと、第2の文字列属性を含む第2のテーブルとを取得してもよいし、予測対象および第1の構造的属性を含む第1のテーブルと、第2の構造的属性を含む第2のテーブルとを取得してもよい。なお、構造的属性については後述される。
- [0024] さらに、入力部10は、第1の属性と第2の属性との類似度を算出するための関数(以下、類似度関数と記す)と、どの程度の類似度の場合に第1の属性の値と第2の属性の値とが類似すると判断するか決定するための条件(

以下、類似度に対する条件と記すこともある)を受け付ける。類似度関数は、数式で表されていてもよく、パラメータとして表されていてもよい。また、類似度に対する条件は、関係の程度に基づいて類似度の有無を判断するための閾値(以下、単に類似度の閾値と記す。)で表されていてもよく、パラメータ等に応じて類似か否かを出力する式で表されていてもよい。

[0025] 例えば、地理的な観点が共通する場合、入力部10は、地理的關係を類似度関数として受け付け、地理的關係の程度を示す類似度の閾値を条件として受け付けてもよい。すなわち、第1の属性および第2の属性が地理的属性である場合、類似度関数は、例えば、距離が近いほど類似度を高く算出する関数として定義される。

[0026] また、例えば、時間的な観点が共通する場合、入力部10は、時間的關係を類似度関数として受け付け、時間的關係の程度を示す類似度の閾値を条件として受け付けてもよい。すなわち、第1の属性および第2の属性が時間的属性である場合、類似度関数は、例えば、時間の差異が小さいほど類似度を高く算出する関数として定義される。

[0027] 他にも、文字列の観点が共通する場合、入力部10は、文字列の關係を類似度関数として受け付け、文字列の關係の程度を示す類似度の閾値を条件として受け付けてもよい。具体的には、第1の属性および第2の属性が文字列属性である場合、類似度関数は、例えば、二つのテキストの一致度が高いほど類似度を高く算出する関数として定義される。テキストの類似度として、例えば、形態素のSimpson係数が挙げられる。

[0028] $morph(a)$ をテキスト文字列 a に含まれる形態素の集合と定義する。例えば、アドレスを示す以下の4つのテキスト文字列は、形態素解析により、それぞれ以下のような形態の集合として表される。

・ $morph('川崎市中原区') = \{'川崎', '市', '中原', '区'\}$

・ $morph('神奈川県川崎市中原区') = \{'神奈川', '県', '川崎', '市', '中原', '区'\}$

・ $morph('神奈川県川崎市幸区') = \{'神奈川', '県', '川崎', '市', '幸', '区'\}$

’区’}

・ $\text{morph}(\text{'神奈川県横浜市港南区'}) = \{\text{'神奈川県'}, \text{'横浜市'}, \text{'港南区'}\}$

[0029] また、テキスト文字列 a とテキスト文字列 b の類似度を算出する関数 $\text{textSim}(a, b)$ は、以下に示す式 1 で定義できる。

[0030]
$$\text{textSim}(a, b) = \frac{|\text{morph}(a) \cup \text{morph}(b)|}{\min(|\text{morph}(a)|, |\text{morph}(b)|)}$$

...

(式 1)

[0031] この場合、上記に例示するアドレスのテキスト文字列同士の類似度は、以下のように算出される。

[0032] ・ $\text{textSim}(\text{'川崎市中原区'}, \text{'神奈川県川崎市中原区'}) = 4 / 4 = 1.0$

・ $\text{textSim}(\text{'川崎市中原区'}, \text{'神奈川県川崎市幸区'}) = 3 / 4 = 0.75$

・ $\text{textSim}(\text{'川崎市中原区'}, \text{'神奈川県横浜市港南区'}) = 2 / 4 = 0.5$

[0033] また、構造的な観点が共通する場合、入力部 10 は、構造的関係を類似度関数として受け付け、構造的関係の程度を示す類似度の閾値を条件として受け付けてもよい。以下、住所やファイルのディレクトリ構造など、木構造の情報を “/” で表現した文字列をパス文字列と定義する。例えば、住所「神奈川県川崎市」は、パス文字列では ‘/神奈川県/川崎市’ と表現される。また、例えば、ディレクトリ構造「news→economy→bigdata」は、パス文字列では、 ‘news/economy/bigdata’ と表現される。

[0034] 第 1 の属性および第 2 の属性が上述するパス文字列で定義される構造的属性の場合、類似度関数は、例えば、二つのパス文字列の距離に近いほど類似

度を高く算出する関数として定義される。パス文字列の距離関数として、例えば、最低共通祖先ノード（LCA: Lowest common ancestor）への距離の最小値が挙げられる。

[0035] 最低共通祖先ノードとは、二つのパスが表現するそれぞれ一番下のノードから上位（先祖）方向に辿った場合に、最初に現れる同じノードである。また、最低共通祖先ノードへの距離とは、一番下のノードから最低共通祖先ノードへ辿ったときのノード数である。

[0036] 例えば、二つのパス文字列 ' $/a/b/c$ ' , ' $/a/b/z$ ' が存在するとする。この場合、二つのパスの最低共通祖先ノードは、 ' $/a/b$ ' である。また、 ' $/a/b/c$ ' から ' $/a/b$ ' への距離は1であり、 ' $/a/b/z$ ' から ' $/a/b$ ' への距離も1である。

[0037] また、例えば、二つのパス文字列 ' $/a/b/c$ ' , ' $/a/d/e/z$ ' が存在するとする。この場合、二つのパスの最低共通祖先ノードは、 ' $/a$ ' である。また、 ' $/a/b/c$ ' から ' $/a$ ' への距離は2であり、 ' $/a/d/e/z$ ' から ' $/a$ ' への距離は3である。

[0038] パス文字列の距離を表す関数を $pathDis(x, y)$ とすると、上述するパス文字列の距離は、以下のように算出される。

[0039] $\cdot pathDis('/a/b/c', '/a/b/z') = 1$
 $\cdot pathDis('/a/b/c', '/a/d/e/z') = 2$

[0040] 図2は、コンフィギュレーションファイル（以下、コンフィグファイルと記す。）の例を示す説明図である。図2に示す例では、類似度関数および類似度に対する条件が、コンフィギュレーションファイル（以下、コンフィグファイルと記す。）に設定されていることを示す。入力部10は、このコンフィグファイルを受け付けてもよい。

[0041] 図2に例示するコンフィグファイルのC1部分が、類似度関数および類似度に対する条件を示す。なお、コンフィグファイルのC2～C4部分については、後述される。C1部分において、前段部（コロンの前）が、第1の属性のデータ型（より具体的には、分析データ型）と第2の属性のデータ型（

より具体的には、分析データ型）との対応関係を示す。また、後段部（コロンの後）が、類似度関数および条件（類似度の閾値）を示す。以下、各内容について、詳細に説明する。

[0042] C1部分における“Point-Point”の行は、点で表される第1の地理的属性と点で表される第2の地理的属性との距離を表す地理的关系を定義する。

[0043] “DistanceMap”は、地理的关系の程度を規定したマップ関数であり、パラメータとして、距離の閾値を含む。DistanceMap関数における3つのパラメータは、順に“開始値”、“終了値”、（開始値から終了値までに適用する閾値の）“間隔”を示す。図2に例示する（“DistanceMap”, 1, 3, 1）は、距離の単位をkmとすると、“距離が1km以内”、“距離が2km以内”、“距離が3km以内”の3つの閾値を関数に適用することを示す。

[0044] “KNearestMap”は、地理的关系の程度を規定したマップ関数であり、パラメータとして、近接する地理的情報の個数の閾値を含む。KNearestMap関数における3つのパラメータも同様、順に“開始値”、“終了値”、（開始値から終了値までに適用する閾値の）“間隔”を示す。図2に例示する（“KNearestMap”, 3, 5, 1）は、近接する地理的情報の個数が、“3つ以内”、“4つ以内”、“5つ以内”の3つの閾値を関数に適用することを示す。

[0045] “SameCityMap”は、地理的关系の程度を規定したマップ関数であり、2つの地点が同じエリアに含まれるか否かを判断する関数である。SameCityMap関数はパラメータを含まないが、エリアを定義したエリア情報に基づいて同じエリアに含まれるか否かが判断される。エリア情報は、予め定義される。

[0046] C1部分における“Point-Area”の行は、点で表される第1の地理的属性と領域で表される第2の地理的属性との包含関係を表す地理的关系を定義する。

- [0047] “InclusionMap”は、地理的關係の程度を規定したマップ関数であり、点で表される第1の地理的属性が領域で表される第2の地理的属性に含まれるか否かを判断する関数である。なお、InclusionMapは、パラメータを含まない。
- [0048] また、“Point-Area”の行においても、“KNear est Map”が定義される。KNear est Map関数の内容は、“Point-Point”におけるKNear est Map関数と同様である。
- [0049] C1部分における“Area-Area”の行は、領域で表される第1の地理的属性と領域で表される第2の地理的属性との交差關係を表す地理的關係を定義する。
- [0050] “IntersectMap”は、地理的關係の程度を規定したマップ関数であり、領域で表される第1の地理的属性が領域で表される第2の地理的属性と交差するか否かを判断する関数である。なお、IntersectMapは、パラメータを含まない。
- [0051] 以上に示すように、第1の地理的データ型と第2の地理的データ型とは、互いに同一の地理的データ型であってもよく、異なる地理的データ型であってもよい。また、第1の地理的データ型が、点の情報で地理を特定可能なデータのタイプであり、第2の地理的データ型が、範囲の情報で地理を特定可能なデータのタイプであってもよい。
- [0052] C1部分における“Time Stamp-Time Stamp”の行は、第1の時間的属性と第2の時間的属性との差異を表す時間的關係を定義する。
- [0053] “TimeDiffMap”は、時間的關係の程度を規定したマップ関数であり、パラメータとして、時間の差異の閾値を含む。TimeDiffMap関数における3つのパラメータも同様、順に“開始値”、“終了値”、（開始値から終了値までに適用する閾値の）“間隔”を示す。図2に例示する（“TimeDiffMap”，30，60，30）は、時間の単位を分とすると、“時間の差異が30分以内”、“時間の差異が60分以内”の2

つの閾値を関数に適用することを示す。

[0054] C 1 部分における “Text-Text” の行は、文字列を表す第 1 の属性と文字列を表す第 2 の属性との一致関係を定義する。“ExactMap” は、文字列で表される属性が一致するか否かを判断する関数である。

[0055] また、“Text-Text” の行に、文字列を表す第 1 の属性と文字列を表す第 2 の属性との類似関係を定義してもよい。具体的には、“Text-Text” の行に、文字列の関係の程度を規定したマップ関数 “textSimMap” を設定してもよい。“textSimMap” は、文字列の関係の程度を規定したマップ関数であり、パラメータとして、類似度の閾値を含む。textSimMap 関数は、DistanceMap 関数同様、3 つのパラメータを有し、それぞれ順に “開始値”、“終了値”、(開始値から終了値までに適用する閾値の) “間隔” を示す。

[0056] 例えば、textSimMap 関数を用いて [(“textSimMap”, 0.8, 1.0, 0.1)] と定義されていたとする。これは、“類似度が 0.8 以上”、“類似度が 0.9 以上” および “類似度が 1.0 (以上)” の 3 つの閾値を関数に適用することを示す。

[0057] なお、類似度関数および類似度の閾値の設定方法は、図 2 の C 1 部分に例示する内容に限定されない。コンフィグファイルに、例えば、パス文字列で表される第 1 の構造的属性とパス文字列で表される第 2 の構造的属性との距離を表す構造的関係 “Path-Path” を定義してもよい。

[0058] 具体的には、“Path-Path” の行に、構造的関係の程度を規定したマップ関数 “pathDisMap” を設定してもよい。“pathDisMap” は、構造的関係の程度を規定したマップ関数であり、パラメータとして、距離の閾値を含む。pathDisMap 関数は、DistanceMap 関数同様、3 つのパラメータを有し、それぞれ順に “開始値”、“終了値”、(開始値から終了値までに適用する閾値の) “間隔” を示す。

[0059] 例えば、pathDisMap 関数を用いて [(“pathDisMap”, 1, 3, 1)] と定義されていたとする。これは、“距離が 1 以下”、“

距離が2以下” および “距離が3以下” の3つの閾値を関数に適用すること
を示す。

[0060] 入力部10が、図2に例示するコンフィグファイルを受け付けることで、
後述するマップパラメータ生成器30が、第1のテーブルに含まれるレコー
ドと第2のテーブルに含まれるレコードとを結合するための結合条件（マッ
プパラメータ）を生成する。

[0061] なお、入力部10は、テーブルの各列が示すデータの属性を合わせて受け
付けてもよい。

[0062] ジオコード20は、文字列で表された属性のデータを変換する。例えば、
地理的属性のデータが文字列で表されている場合、ジオコード20は、その
文字列を、点、多角形または複数多角形のデータに変換する。なお、データ
を変換する必要がない場合、情報処理システム100は、ジオコード20を
備えていなくてもよい。

[0063] 図3は、データを変換する処理の例を示す説明図である。図3に示す例で
は、列ごとの分析データ型を定義したテーブルadt1と、分析データ型か
らデータ型へ変換する対応を定義したテーブルadt2が予め取得されてい
るものとする。

[0064] この状況で、入力部10が、図3に例示するターゲットテーブルT、ソー
ステーブルS1およびソーステーブルS2を取得したとする。ソーステーブ
ルS2の“Pickup_location”列の分析データ型は、テーブ
ルadt1を参照するとPointであり、変換の必要がない。一方、ソー
ステーブルS1の“community”列の分析データ型はテーブルadt
1を参照すると“TownAddress”であり、テーブルadt2を
参照すると、データ型Polygonに変換する必要がある。そこで、ジオ
コード20は、ソーステーブルS1の“community”列に含まれる
データを、多角形の領域で表すように変換する。例えば、“communi
ty”の内容に応じて多角形で領域を特定可能なエリア情報を予め定めてお
き、ジオコード20は、そのエリア情報に基づいて、データ型がPolyg

onになるようにデータを変換してもよい。

- [0065] マップパラメータ生成器30、フィルタパラメータ生成器50、および、集約パラメータ生成器60は、後述する特徴量生成関数生成器81が、予測対象に影響を及ぼし得る変数である特徴量を生成するための特徴量生成関数を生成する際に利用するパラメータを生成する。
- [0066] 以下の説明では、特徴量とは、特徴そのものの内容（例えば、「人口」、「位置」など）を意味する。また、特徴量に具体的なデータをあてはめたもの（例えば、人口＝“8112”、位置＝“(−73.965, 40.724)”など）のことを、特徴量ベクトル（複数の場合、特徴量テーブル）と記す。
- [0067] また、後述する特徴量生成器82が生成する特徴量は、機械学習を用いてモデルを生成する際の説明変数の候補になる。言い換えると、本実施形態で生成される特徴量生成関数を用いることで、機械学習を用いてモデルを生成する際の説明変数の候補を自動的に生成することが可能になる。
- [0068] 図4は、各パラメータと、第1のテーブルおよび第2のテーブルとの関係の例を示す説明図である。
- [0069] フィルタパラメータ生成器50が生成するパラメータは、第2のテーブルに含まれる行の抽出条件を表わすパラメータである。以下、このパラメータをフィルタパラメータとし、フィルタパラメータに基づいて第2のテーブルから行を抽出する処理を「filter」と記載する場合がある。また、この抽出条件のリストを「Fリスト」と記載する場合がある。抽出条件は任意であり、例えば、指定された列の値と同じ（大きいまたは小さい）か否か判断する条件が挙げられる。
- [0070] 集約パラメータ生成器60が生成するパラメータは、第2のテーブルに含まれる各行のデータを目的変数ごとに集約する集約方法を表わすパラメータである。なお、一般に、第1の表における行と第2の表における行とは、一対多対応する場合が多いため、結果として行が集約されることになる。集約情報は、ソーステーブル（第2のテーブル）の列に対する集約関数として定

義されてもよい。

[0071] 集約方法は任意であり、例えば、列の総数、最大値、最小値、平均値、中央値、分散などが挙げられる。また、列の総数の集計は、重複データを除外する、または、重複データを除外しない、のいずれかの観点で行われてもよい。

[0072] 以下、このパラメータを集約パラメータと記し、集約パラメータが示す方法により各列のデータを集約する処理を「`reduce`」と記載する場合がある。特に、地理的情報を集約する処理を「`Geo-reduce`」と記載することもある。また、この集約処理のリストを「`R`リスト」と記載する場合がある。なお、地理的情報を集約する処理の詳細については後述される。

[0073] マップパラメータ生成器30が生成するパラメータは、第1のテーブルと第2のテーブルの列との対応条件を表わすパラメータである。以下、このパラメータをマップパラメータと記し、マップパラメータに基づいて各テーブルの列を対応付ける処理を「`map`」と記載する場合がある。また、この対応条件のリストを「`M`リスト」と記載する場合がある。特に、地理的情報同士を対応付ける処理を「`Geo-map`」と記載することもある。また、`map`による各テーブルの列の対応付けは、対応付けられた列で複数の表を1つの表に結合（`join`）することとも言える。なお、地理的情報を対応付ける処理の詳細についても後述される。

[0074] マップパラメータ生成器30は、ジオマップ生成器（GeoMap Generator）40と、時間差異マップ生成器（TimeDiff Map Generator）31と、マップ生成器（Exact Map Generator）32と、属性特定部33とを含む。マップパラメータ生成器30（より具体的には、マップパラメータ生成器30に含まれる各生成器）は、第1の属性の値と第2の属性の値とにより算出される類似度が、条件を満たすような第1の属性の値を含む第1のテーブルのレコードと、第2の属性の値を含む第2のテーブルのレコードとを結合するための結合条件を生成する。条件を満たすとは、例えば、類似度が閾値以下または以上になることや、予め定めた範囲内に含まれることなどを意味する。

- [0075] ジオマップ生成器40は、第1のテーブルと第2のテーブルの地理的属性を含む列同士に対応条件を表すパラメータを生成する。ジオマップ生成器40は、距離マップ生成器(Distance Map Generator)41と、包含マップ生成器(Inclusion Map Generator)42と、重複マップ生成器(Overlap Map Generator)43と、同地域マップ生成器(SameArea Map Generator)44とを有する。
- [0076] ジオマップ生成器40(より具体的には、ジオマップ生成器40に含まれる各生成器)は、第1の地理的属性の値と第2の地理的属性の値との関係が、地理的關係の程度を満たすような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するための結合条件(マップパラメータ)を生成する。以下、各生成器の処理について、詳細に説明する。
- [0077] 距離マップ生成器41は、距離の近さに基づいて第1のテーブルと第2のテーブルを対応付けるための類似度関数および条件(例えば、類似度の閾値)を受け付けた場合に、マップパラメータを生成する。図2に示す例では、DistanceMap関数とKNearestMap関数の少なくとも一方がコンフィグファイルに設定されている場合に対応する。
- [0078] 距離マップ生成器41は、第1の地理的属性の値と第2の地理的属性の値との距離が閾値以内であるような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するためのマップパラメータを生成する。
- [0079] 図5は、距離に基づいてマップパラメータを生成する処理の例を示す説明図である。図5に示す例では、ターゲットテーブルTとソーステーブルS2をそれぞれ1つずつ取得した場合を示す。なお、図5に例示するターゲットテーブルTは、2015年1月8日22時の、5か所における乗客数(pickup_number)を表すデータを含むテーブルである。また、図5に例示するソーステーブルS2は、各時刻における乗客数、移動距離および乗客の乗り場位置を対応付けて記録するテーブルである。
- [0080] 例えば、図2に例示するDistanceMap関数の場合、距離マップ

生成器41は、第1の地理的属性の値が示す位置と第2の地理的属性の値が示す位置との距離が1km以内であるソーステーブルS2のレコードにターゲットテーブルTの各レコードを対応付けるパラメータを生成する。さらに、距離マップ生成器41は、第1の地理的属性の値が示す位置と第2の地理的属性の値が示す位置との距離が2km以内および3km以内であるソーステーブルS2のレコードにターゲットテーブルTの各レコードを対応付けるパラメータをそれぞれ生成する。

[0081] 図5に示す例では、ターゲットテーブルTの“target_location”列の属性が第1の地理的属性であり、ソーステーブルS2の“Pickup_location”列の属性が第2の地理的属性である。この2つの列が対応付けられる。なお、第1のテーブルと第2のテーブルとで対応付ける列は、予め指定されていてもよく、後述する属性特定部33によって特定されてもよい。

[0082] この結果、図5に例示するパラメータP11が生成される。図5に例示するように、地理的分析データ型に基づいてマップパラメータが生成され、1つのマップパラメータに基づいて、1つのマップ処理が定義される。図5に例示するマップデータM11は、距離が1km以内であるソーステーブルS2のレコードにターゲットテーブルTの各レコードを対応付けた結果を示す。例えば、ターゲットテーブルの1番目のレコードに対して、ソーステーブルから1つのレコードのみ対応付けられる。また、例えば、ターゲットテーブルの2番目のレコードに対して、ソーステーブルから2つのレコードが対応付けられる。

[0083] 図6は、距離に基づいてマップパラメータを生成する他の処理の例を示す説明図である。図6に例示するターゲットテーブルTおよびソーステーブルS2は、図5に例示するターゲットテーブルTおよびソーステーブルS2と同様である。

[0084] 例えば、図2に例示するKNearestMap関数の場合、距離マップ生成器41は、第1の地理的属性の値が示す位置と第2の地理的属性の値が

示す位置との距離が近い方から順にソーステーブルS 2のレコードにターゲットテーブルTの各レコードを2つ以内で対応付けるパラメータを生成する。さらに、距離マップ生成器4 1は、第1の地理的属性の値が示す位置と第2の地理的属性の値が示す位置との距離が近い方から順にソーステーブルS 2のレコードにターゲットテーブルTの各レコードを3つ以内および4つ以内で対応付けるパラメータをそれぞれ生成する。

[0085] 図6に示す例では、ターゲットテーブルTの“`target_location`”列の属性が第1の地理的属性であり、ソーステーブルS 2の“`Pickup_location`”列の属性が第2の地理的属性である。この2つの列が対応付けられる。なお、第1のテーブルと第2のテーブルとで対応付ける列は、予め指定されていてもよく、後述する属性特定部3 3によって特定されてもよい。

[0086] この結果、図6に例示するパラメータP 1 2が生成される。図6に例示するように、地理的分析データ型に基づいてマップパラメータが生成され、1つのマップパラメータに基づいて、1つのマップ処理が定義される。図6に例示するマップデータM 1 2は、近い順にソーステーブルS 2のレコードにターゲットテーブルTの各レコードを2つ対応付けた結果を示す。例えば、ターゲットテーブルの各レコードに対して、ソーステーブルから2つの最も近いレコードが対応付けられる。

[0087] 同地域マップ生成器4 4は、同じエリアに含まれるか否かに基づいて第1のテーブルと第2のテーブルを対応付けるための類似度関数を受け付けた場合に、マップパラメータを生成する。図2に示す例では、`SameCityMap`関数がコンフィグファイルに設定されている場合に対応する。

[0088] 同地域マップ生成器4 4は、第1の地理的属性の値が示す位置と第2の地理的属性の値が示す位置が同じエリアに含まれるような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するためのマップパラメータを生成する。

[0089] 図7は、同じエリアに含まれているか否か判断する方法の例を示す説明図

である。図7に示す例では、各エリアと多角形で特定されるエリアの領域とが対応付けられた共通エリアテーブルC A Tが予め定義される。共通エリアの例として、国、州、都市、自治区、街などが挙げられる。共通エリアは、互いに重なり合わない共通の領域として定義され、マップ上の境界情報を表す。共通エリアテーブルC A Tは、例えば、記憶部80に記憶されていてもよい。

[0090] まず、共通エリアテーブルC A Tに基づいて2つの位置が同じエリアに存在するか否かが判断される。具体的には、ターゲットテーブルTのレコードt 1の位置が示すエリアが特定され、ソーステーブルSのレコードs 1の位置がそのエリア内か否かが判断される。以下、同様の処理が、ターゲットテーブルTおよびソーステーブルSの全てのレコードに対して行われる。

[0091] 図8は、共通エリアか否かに基づいてマップパラメータを生成する処理の例を示す説明図である。図8に例示するターゲットテーブルTおよびソーステーブルS 2は、図5に例示するターゲットテーブルTおよびソーステーブルS 2と同様である。

[0092] 例えば、図2に例示するSameCityMap関数の場合、同地域マップ生成器44は、第1の地理的属性の値が示す位置と第2の地理的属性の値が示す位置とが同じエリアに含まれるソーステーブルS 2のレコードとターゲットテーブルTの各レコードとを対応付けるパラメータを生成する。

[0093] 図8に示す例では、ターゲットテーブルTの“target_location”列の属性が第1の地理的属性であり、ソーステーブルS 2の“Pickup_location”列の属性が第2の地理的属性である。この2つの列が対応付けられる。なお、第1のテーブルと第2のテーブルとで対応付ける列は、予め指定されていてもよく、後述する属性特定部33によって特定されてもよい。

[0094] この結果、図8に例示するパラメータP 13が生成される。図8に例示するマップデータM 13は、同じエリアと判断された地理的属性を有するソーステーブルS 2のレコードとターゲットテーブルTの各レコードとを対応付

けた結果を示す。なお、図8に例示するマップデータM13は、距離が1 km未満の地点が同じ都市に位置すると仮定して対応付けた例を示す。

[0095] 包含マップ生成器42は、包含関係に基づいて第1のテーブルと第2のテーブルを対応付けるための類似度関数を受け付けた場合に、マップパラメータを生成する。図2に示す例では、`InclusionMap`関数がコンフィグファイルに設定されている場合に対応する。

[0096] 包含マップ生成器42は、第1の地理的属性の値が示す位置が第2の地理的属性の値が示す領域に含まれているような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するためのマップパラメータを生成する。

[0097] 図9は、包含関係に基づいてマップパラメータを生成する処理の例を示す説明図である。図9に例示するターゲットテーブルTは、図5に例示するターゲットテーブルTと同様である。また、図9に例示するソーステーブルS1は、各領域における人口、男性数および20歳から40歳までの人数を対応付けて記録するテーブルである。

[0098] 例えば、図2に例示する`InclusionMap`関数の場合、包含マップ生成器42は、第1の地理的属性の値が示す位置が第2の地理的属性の値が示す領域に含まれるソーステーブルS1のレコードにターゲットテーブルTの各レコードを対応付けるパラメータを生成する。

[0099] 図9に示す例では、ターゲットテーブルTの“`target_location`”列の属性が第1の地理的属性であり、ソーステーブルS1の“`community`”列の属性が第2の地理的属性である。この2つの列が対応付けられる。なお、第1のテーブルと第2のテーブルとで対応付ける列は、予め指定されていてもよく、後述する属性特定部33によって特定されてもよい。

[0100] この結果、図9に例示するパラメータP14が生成される。図9に例示するマップデータM14は、同じエリアに存在するソーステーブルS1のレコードにターゲットテーブルの各レコードを対応付けた結果を示す。

- [0101] 重複マップ生成器43は、重複する領域に基づいて第1のテーブルと第2のテーブルを対応付けるための類似度関数を受け付けた場合に、マップパラメータを生成する。図2に示す例では、`IntersectMap`関数がコンフィグファイルに設定されている場合に対応する。
- [0102] 重複マップ生成器43は、第1の地理的属性の値が示す領域と第2の地理的属性の値が示す領域が重複するような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するためのマップパラメータを生成する。
- [0103] 時間差異マップ生成器31は、時間の差異に基づいて第1のテーブルと第2のテーブルを対応付けるための類似度関数および条件（例えば、類似度の閾値）を受け付けた場合に、マップパラメータを生成する。図2に示す例では、`TimeDiffMap`関数がコンフィグファイルに設定されている場合に対応する。
- [0104] 時間差異マップ生成器31は、第1の時間的属性の値と第2の時間的属性の値との関係が時間的關係の程度を満たすような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するための結合条件を生成する。本実施形態では、時間差異マップ生成器31は、第1の時間的属性の値と第2の時間的属性の値との差異が閾値以内であるような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するためのマップパラメータを生成する。
- [0105] 図10は、時間の差異に基づいてマップパラメータを生成する処理の例を示す説明図である。図10に例示するターゲットテーブルTおよびソーステーブルS2は、図5に例示するターゲットテーブルTおよびソーステーブルS2と同様である。
- [0106] 例えば、図2に例示する`TimeDiffMap`関数の場合、時間差異マップ生成器31は、第1の時間的属性の値と第2の地理的属性の値との差異が30分以内であるソーステーブルS2のレコードにターゲットテーブルTの各レコードを対応付けるパラメータを生成する。さらに、時間差異マップ

生成器 3 1 は、第 1 の時間的属性の値と第 2 の時間的属性の値との差異が 60 分以内であるソーステーブル S 2 のレコードにターゲットテーブル T の各レコードを対応付けるパラメータを生成する。

[0107] 図 10 に示す例では、ターゲットテーブル T の “time” 列の属性が第 1 の時間的属性であり、ソーステーブル S 2 の “pickup_time” 列の属性が第 2 の時間的属性である。この 2 つの列が対応付けられる。なお、第 1 のテーブルと第 2 のテーブルとで対応付ける列は、予め指定されていてもよく、後述する属性特定部 33 によって特定されてもよい。

[0108] この結果、図 10 に例示するパラメータ P 15 が生成される。図 10 に例示するマップデータ M 15 は、時間の差異が 30 分以内であるソーステーブル S 2 のレコードにターゲットテーブル T の各レコードを対応付けた結果を示す。

[0109] マップ生成器 32 は、第 1 のテーブルと第 2 のテーブルを対応付けるための類似度関数を受け付けた場合に、マップパラメータを生成する。本実施形態では、地理的属性と時間的属性のいずれの属性でもない属性の値に基づいてソーステーブルのレコードにターゲットテーブルのレコードを対応付けるパラメータを生成する。

[0110] 図 2 に示す例では、ExactMap 関数がコンフィグファイルに設定されている場合に対応する。マップ生成器 32 は、第 1 の属性の値と第 2 の属性の値とが一致するような、第 1 のテーブルに含まれるレコードと第 2 のテーブルに含まれるレコードとを結合するためのマップパラメータを生成する。

[0111] 図 11 は、テキストの類似性に基づいてマップパラメータを生成する処理の例を示す説明図である。図 11 に例示するターゲットテーブル T は、ある住所における乗客数 (pickup_number) を表すデータを含むテーブルである。また、図 11 に例示するソーステーブル S は、各地域における収入平均を記録するテーブルである。

[0112] 例えば、上述する textSimMap 関数の場合、マップ生成器 32 は

、第1の文字列属性の値と第2の文字列属性の値との類似度が0.8以上であるソーステーブルSのレコードにターゲットテーブルTの各レコードを対応付けるパラメータを生成する。さらに、マップ生成器32は、第1の文字列属性の値と第2の文字列属性の値との類似度が0.9以上および1.0以上であるソーステーブルSのレコードにターゲットテーブルTの各レコードを対応付けるパラメータをそれぞれ生成する。

[0113] 図11に示す例では、ターゲットテーブルTの“address”列の属性が第1の文字列属性として、ソーステーブルSの“address”列の属性が第2の文字列属性として、それぞれ登録されているとする。そこで、この2つの列が対応付けられる。この結果、図11に例示するパラメータP16が生成される。

[0114] 図11に例示するマップデータMは、類似度が0.8以上であるソーステーブルSのレコードにターゲットテーブルTの各レコードを対応付けた結果を示す。例えば、ターゲットテーブルの1番目のレコードに対して、ソーステーブルから1つのレコードのみ対応付けられる。

[0115] 図12は、構造の類似性に基づいてマップパラメータを生成する処理の例を示す説明図である。図12に例示するターゲットテーブルTは、あるURLで識別されるWebページへのアクセス数(access_number)を表すデータを含むテーブルである。また、図12に例示するソーステーブルSは、あるURLで識別されるWebページの先月のアクセス数(access_number)を記録するテーブルである。

[0116] 例えば、上述するpathDisMap関数の場合、マップ生成器32は、第1の構造的属性の値と第2の構造的属性の値との距離が1以下であるソーステーブルSのレコードにターゲットテーブルTの各レコードを対応付けるパラメータを生成する。さらに、マップ生成器32は、第1の構造的属性の値と第2の構造的属性の値との距離が2以下および3以下であるソーステーブルSのレコードにターゲットテーブルTの各レコードを対応付けるパラメータをそれぞれ生成する。

- [0117] 図 1 2 に示す例では、ターゲットテーブル T の “U R L” 列の属性が第 1 の構造的属性として、ソーステーブル S の “U R L” 列の属性が第 2 の構造的属性として、それぞれ登録されているとする。そこで、この 2 つの列が対応付けられる。この結果、図 1 2 に例示するパラメータ P 1 7 が生成される。
- [0118] 図 1 2 に例示するマップデータ M は、類似度が 1 以下であるソーステーブル S のレコードにターゲットテーブル T の各レコードを対応付けた結果を示す。例えば、ターゲットテーブルの 1 番目のレコードに対して、ソーステーブルから 1 つのレコードのみ対応付けられる。
- [0119] 属性特定部 3 3 は、第 1 のテーブルと第 2 のテーブルとで、観点が共通する属性を特定する。具体的には、属性特定部 3 3 は、第 1 のテーブルの各列が示すデータの属性と、第 2 のテーブルの各列が示すデータの属性とが同じ属性を特定する。例えば、地理的データ型の場合、属性特定部 3 3 は、第 1 の地理的データ型と同じデータ型を有する第 1 の地理的属性を第 1 のテーブルから特定し、且つ、第 2 の地理的情報のデータ型と同じデータ型を有する第 2 の地理的属性を第 2 のテーブルから特定してもよい。このようにすることで、地理的データ型を有する列を各テーブルから特定することが可能になる。また、属性特定部 3 3 は、入力部 1 0 に入力された列の属性の情報から、第 1 のテーブルと第 2 のテーブルの列の属性を特定してもよい。
- [0120] マップパラメータ生成器 3 0（より具体的には、マップパラメータ生成器 3 0 に含まれる各生成器）は、地理的（時間的）関係の判断対象である第 1 の地理的（時間的）属性を含む第 1 のテーブルの列および第 2 の地理的（時間的）属性を含む第 2 のテーブルの列と、地理的（時間的）関係の程度とを含むパラメータを、記憶部 8 0 に記憶させてもよい。例えば、マップパラメータ生成器 3 0 は、図 5 に例示するパラメータ P 1 1 や図 1 0 に例示するパラメータ P 1 5 などを記憶部 8 0 に記憶させてもよい。
- [0121] 図 1 3 は、生成されたマップパラメータの例を示す説明図である。上述する例で示すように、入力部 1 0 が、図 1 3 に例示するターゲットテーブル T

、ソーステーブルS 1およびソーステーブルS 2、並びに、図2に例示するコンフィグファイルのC 1部分を受け付ける。なお、マップパラメータP 16は、ターゲットテーブルTの“t a r g e t _ l o c a t i o n”列の属性を第1の地理的属性とし、ソーステーブルS 1の“c o m m u n i t y”列の属性を第2の地理的属性として、K N e a r e s t M a p関数に基づいて生成されるパラメータの例である。マップパラメータ生成器30（より具体的には、マップパラメータ生成器30に含まれる各生成器）は、これらの情報から、図13に例示する13個のマップパラメータP 11～16を生成する。

[0122] フィルタパラメータ生成器50は、フィルタ生成器（Exact Filter Generator）51を含む。フィルタ生成器51は、第2のテーブルの列と、その列に適用する抽出条件とを対応付けたフィルタパラメータを生成する。

[0123] フィルタパラメータの生成方法は任意である。フィルタ生成器51は、例えば、図2に例示するコンフィグファイルのC 2部分で定義された情報に基づいて、フィルタパラメータを生成してもよい。また、予め記憶部80に抽出条件を記憶しておき、フィルタ生成器51は、その抽出条件を読み取ってフィルタパラメータを生成してもよい。

[0124] さらに、フィルタ生成器51は、抽出条件を複数組み合わせ、さらなる抽出条件を生成してもよい。また、抽出条件を組み合わせる数も任意である。入力部10は、例えば、この組み合わせ最大数を受け付けてもよい。例えば、図2に例示するように、コンフィグファイルのC 4部分に組み合わせ最大数を示すパラメータ（“m a x _ c o m b i n a t i o n _ f i l t e r _ l e n g t h”）が設定されていてもよい。

[0125] 集約パラメータ生成器60（より具体的には、集約パラメータ生成器60に含まれる各生成器）は、第2のテーブルに含まれる各行のデータを集約する方法を表わすパラメータを生成する。集約パラメータ生成器60は、ジオ集約生成器（GeoReduce Generator）70と、数的集約生成器（Numeric Reduce Generator）61とを含む。

- [0126] ジオ集約生成器 70（より具体的には、ジオ集約生成器 70に含まれる各生成器）は、第2のテーブルに含まれる地理的属性を含む列の値で各行のデータを集約する方法を表わす集約パラメータを生成する。具体的には、ジオ集約生成器 70は、指定された集約方法に基づいて地理的属性の値の統計値を算出する。
- [0127] 集約方法を指定する方法は任意である。例えば、入力部 10が集約方法の指定を受け付けてもよい。具体的には、図2のコンフィグファイルのC3部分に例示するように、地理的属性の分析データ型に応じて集約方法を定義し、定義された集約方法に応じて集約パラメータを生成してもよい。以下、各内容について、詳細に説明する。
- [0128] C3部分における“Point”の行は、第2の地理的属性（より具体的には、地理的データ型）が点（Point）で表される場合の集約方法を定義する。
- [0129] （“sum”，“distance”）は、第1のテーブルのレコードに対応付けられた第2のテーブルの各レコードのうち、第1の地理的属性の値と第2の地理的属性の値とに基づいて算出される距離の合計を統計値として算出する集約方法を定義する。
- [0130] （“avg”，“distance”）は、第1のテーブルのレコードに対応付けられた第2のテーブルの各レコードのうち、第1の地理的属性の値と第2の地理的属性の値とに基づいて算出される距離の平均を統計値として算出する集約方法を定義する。
- [0131] （“count”）は、第1のテーブルの各レコード（すなわち、目的変数）に対応付けられた第2のテーブルのレコード数を統計値として算出する集約方法を定義する。
- [0132] C3部分における“Area”の行は、第2の地理的属性（より具体的には、地理的データ型）が領域（Area）で表される場合の集約方法を定義する。
- [0133] （“sum”，“areaSize”）は、第1のテーブルのレコードに

対応付けられた第2のテーブルの各レコードのうち、第2の地理的属性の領域の大きさの合計を統計値として算出する集約方法を定義する。

[0134] (“avg”, “areaSize”) は、第1のテーブルのレコードに対応付けられた第2のテーブルの各レコードのうち、第2の地理的属性の領域の大きさの平均を統計値として算出する集約方法を定義する。

[0135] (“count”) は、第1のテーブルの各レコード（すなわち、目的変数）に対応付けられた第2のテーブルのレコード数を統計値として算出する集約方法を定義する。

[0136] ジオ集約生成器70は、ポイント集約生成器(Point Reduce Generator)71と、エリア集約生成器(Area Reduce Generator)72とを有する。

[0137] ポイント集約生成器71は、第1の地理的属性の値と第2の地理的属性の値とに基づいて算出される距離の統計値を算出するための集約パラメータを生成する。なお、ここで対象とする第2のテーブルのレコードは、第1のテーブルのレコードにそれぞれ対応付けられたレコードである。地理的属性の場合、上述するように、第1の地理的属性の値と第2の地理的属性の値とが、一致する、または、一定の範囲内にあるなど、一定の条件を満たすレコード同士が対応付けられる。そこで、ポイント集約生成器71は、第1の地理的属性の値に対する第2の地理的属性の値が所定の条件を満たす場合に、第1の地理的属性の値と条件を満たす第2の地理的属性の値とに基づいて距離の統計値を算出するための集約パラメータを生成する。なお、算出される統計値は、特徴量として利用される。

[0138] ポイント集約生成器71は、例えば、図2に例示する(“sum”, “distance”)、(“avg”, “distance”)および(“count”)の少なくとも一つがコンフィグファイルに設定されている場合に、距離の統計値を算出するための集約パラメータを生成してもよい。

[0139] 図14は、距離の統計値を算出するための集約パラメータを生成する処理の例を示す説明図である。図14に示す例では、3種類の集約方法がコンフィグファイルに設定されている。そこで、ポイント集約生成器71は、ソー

ステーブルのレコードとターゲットテーブルのレコードとの間の距離の合計および平均を算出する集約パラメータ、並びに、対応付けられたソーステーブルのレコード数を算出する集約パラメータを算出する。ポイント集約生成器 7 1 は、例えば、図 1 4 に例示する集約リスト P 2 1 のように、集約するソーステーブルの列名、対応付けるターゲットテーブルの列名、集約内容（距離）および集約関数を対応付けた集約パラメータを生成してもよい。

[0140] 図 1 4 に例示する集約データ R 2 1 は、距離の合計を算出する集約パラメータに基づいて、マップデータ M 1 1 を集約した結果を示す。

[0141] エリア集約生成器 7 2 は、第 2 の地理的属性の値に基づいて算出される領域の統計値を算出するための集約パラメータを生成する。なお、ポイント集約生成器 7 1 と同様、ここで対象とする第 2 のテーブルのレコードは、第 1 のテーブルのレコードにそれぞれ対応付けられたレコードである。

[0142] エリア集約生成器 7 2 は、例えば、図 2 に例示する（“sum”，“areaSize”）および（“avg”，“areaSize”）および（“count”）の少なくとも一つがコンフィグファイルに設定されている場合に、領域の統計値を算出するための集約パラメータを生成してもよい。

[0143] 図 1 5 は、領域の統計値を算出するための集約パラメータを生成する処理の例を示す説明図である。図 1 5 に示す例では、3 種類の集約方法がコンフィグファイルに設定されている。そこで、エリア集約生成器 7 2 は、ターゲットテーブルの各レコードに対応付けられたソーステーブルのレコードの面積の合計および平均を算出する集約パラメータ、並びに、対応付けられたソーステーブルのレコード数を算出する集約パラメータを算出する。エリア集約生成器 7 2 は、例えば、図 1 5 に例示する集約リスト P 2 2 のように、集約するソーステーブルの列名、集約内容（面積）および集約関数を対応付けた集約パラメータを生成してもよい。

[0144] 図 1 5 に例示する集約データ R 2 2 は、面積の合計を算出する集約パラメータに基づいて、マップデータ M 1 4 を集約した結果を示す。

[0145] 数的集約生成器 6 1 は、第 2 のテーブルに含まれる数値（N e m u r i c

）の属性（以下、数値属性と記す。）を含む列の値で各行のデータを集約する方法を表わす集約パラメータを生成する。具体的には、数的集約生成器 61 は、指定された集約方法に基づいて数値の統計値を算出する。

[0146] 集約方法を指定する方法は任意である。ジオ集約生成器 70 と同様、例えば、入力部 10 が集約方法の指定を受け付けてもよい。具体的には、図 2 のコンフィグファイルの C 3 部分に例示するように、数値属性に対する集約方法を定義し、定義された集約方法に応じて集約パラメータを生成してもよい。図 2 に示す例では、数値属性の列の合計および平均を算出する集約パラメータを生成するための指定がされている。

[0147] 集約パラメータ生成器 60（より具体的には、集約パラメータ生成器 60 に含まれる各生成器）は、生成した集約パラメータを記憶部 80 に記憶させてもよい。図 16 は、生成された集約パラメータの例を示す説明図である。上述する例で示すように、入力部 10 が、図 16 に例示するターゲットテーブル T、ソーステーブル S1 およびソーステーブル S2、並びに、図 2 に例示するコンフィグファイルの C 3 部分を受け付ける。

[0148] なお、集約パラメータ P23 は、ソーステーブル S2 の数値的属性の列に対する集約パラメータの例である。また、集約パラメータ P24 は、ソーステーブル S1 の数値的属性の列に対する集約パラメータの例である。集約パラメータ生成器 60（より具体的には、集約パラメータ生成器 60 に含まれる各生成器）は、これらの情報から、図 16 に例示する 16 個のマップパラメータ P21～24 を生成する。

[0149] 特徴量生成関数生成器 81 は、第 1 のテーブルおよび第 2 のテーブルから、上述する特徴量を生成するための特徴量生成関数を生成する。具体的には、特徴量生成関数生成器 81 は、上述する結合条件（マップパラメータ）と、集約条件（集約パラメータ）とを用いて（組み合わせて）特徴量生成関数を生成する。また、特徴量生成関数生成器 81 は、結合条件および集約条件に加え、抽出条件（フィルタパラメータ）を用いて（組み合わせて）特徴量生成関数を生成してもよい。

- [0150] また、本実施形態では、特徴量生成関数生成器 81 は、結合条件（マップパラメータ）のうち、地理的属性を対象にしたマップパラメータと時間的属性を対象にしたマップパラメータとを予め結合したマップパラメータを生成してもよい。特徴量生成関数生成器 81 は、例えば、図 2 に例示するコンフィグファイルの C4 部分に示すようなパラメータ “time_spatial_map_combination” に “True” が設定されている場合に、地理的属性を対象にしたマップパラメータと時間的属性を対象にしたマップパラメータとを結合すると判断してもよい。
- [0151] 図 17 は、マップパラメータ同士を結合した例を示す説明図である。例えば、地理的属性を対象にした 6 つのマップパラメータ P11, P12 と、時間的属性を対象にした 2 つのマップパラメータ P15 が存在するとする。このとき、特徴量生成関数生成器 81 は、地理的属性を対象にしたマップパラメータと時間的属性を対象にしたマップパラメータとを 1 つずつ組み合わせて、新たなマップパラメータ P31 を生成してもよい。図 17 に示す例の場合、新しく $6 \times 2 = 12$ のマップパラメータが生成される。
- [0152] 以下、特徴量生成関数生成器 81 が特徴量生成関数を生成する手順を具体的に説明する。ここでは、図 13 に例示するターゲットテーブル T、ソーステーブル S1, S2 が入力されるものとする。また、予測対象の変数（目的変数）は、ターゲットテーブル T に含まれる乗客数（pickup_number）を表す変数である。
- [0153] 図 18 は、パラメータを組み合わせて特徴量生成関数を生成する方法の例を示す説明図である。図 18 (a) は、ターゲットテーブル T とソーステーブル S1 とから特徴量を生成するための特徴量生成関数を生成する組合せ例を示す。また、図 18 (b) は、ターゲットテーブル T とソーステーブル S2 とから特徴量を生成するための特徴量生成関数を生成する組合せ例を示す。なお、図 18 (b) に示す例では、地理的属性を対象にしたマップパラメータと時間的属性を対象にしたマップパラメータとが結合されたマップパラメータが利用されるものとする。

- [0154] 図18(a)に示す例では、4のマッピングパラメータと9の集約パラメータが生成されている。特徴量生成関数生成器81は、これらのマッピングパラメータおよび集約パラメータから、それぞれ1つずつパラメータを選択し、各パラメータの組合せを生成する。この例の場合、各パラメータに基づいて、 $4 \times 9 = 36$ 通りの組合せが生成される。なお、フィルタパラメータが生成されている場合、特徴量生成関数生成器81は、マッピングパラメータ、フィルタパラメータおよび集約パラメータから、それぞれ1つずつパラメータを選択し、各パラメータの組合せを生成する。
- [0155] 図18(b)に示す例でも同様に、14のマッピングパラメータと7の集約パラメータが生成されている。特徴量生成関数生成器81は、これらのマッピングパラメータおよび集約パラメータから、それぞれ1つずつパラメータを選択し、各パラメータの組合せを生成する。この例の場合、各パラメータに基づいて、 $14 \times 7 = 98$ 通りの組合せが生成される。以上より、全部で、 $36 + 98 = 134$ のパラメータの組合せが生成される。
- [0156] 次に、特徴量生成関数生成器81は、生成された組合せに基づいて特徴量生成関数を生成する。具体的には、特徴量生成関数生成器81は、生成された組合せに含まれるパラメータを、表データの操作や定義を行う問合せ言語の形式に変換する。特徴量生成関数生成器81は、例えば、問合せ言語としてSQLを用いてもよい。
- [0157] このとき、特徴量生成関数生成器81は、SQL文を生成するテンプレートに各パラメータを適用して特徴量生成関数を生成してもよい。具体的には、各パラメータを当てはめてSQL文を生成するためのテンプレートを予め用意しておき、特徴量生成関数生成器81は、生成された組合せに含まれる各パラメータを順次テンプレートに適用してSQL文を生成してもよい。この場合、特徴量生成関数は、SQL文として定義され、選択される各パラメータが、SQL文を生成するパラメータに対応する。
- [0158] これらのパラメータの組合せで特徴量を定義すると、多数の種類の特徴量生成関数を単純な要素の組合せとして表現することが可能になる。したがっ

て、複数の表データを利用して効率よく多数の特徴量の候補を生成できる。例えば、上述する例の場合、4つのマップパラメータと9つの集約パラメータ、および、14のマップパラメータと7つの集約パラメータを生成するだけで、130種類の特徴量を容易に生成することが可能になる。また、一度生成した各パラメータの定義は再利用できるため、特徴量生成関数を生成する工数自体も削減できるという効果も得られる。

[0159] 特徴量生成器82は、特徴量生成関数を用いて特徴量を生成する。例えば、特徴量生成関数に、上述する距離の統計値を算出するパラメータが含まれているとする。この場合、特徴量生成器82は、特徴量生成関数に基づいて、第1の地理的属性のレコードごとに、所定の条件を満たす第2のテーブルのレコードを集約する演算を行うことにより、距離の統計値を算出してもよい。

[0160] 具体的には、特徴量生成器82は、第2のテーブルのレコードを集約する演算として、第1の地理的属性の各レコードに対して所定の条件を満たす第2のテーブルの地理的属性との距離の合計と平均の少なくともいずれかを算出してもよい。そして、特徴量生成器82は、算出した距離の合計と平均の少なくともいずれかを特徴量として第1のテーブルの属性に追加してもよい。

[0161] 他にも、特徴量生成器82は、第2のテーブルのレコードを集約する演算として、第1の地理的属性の各レコードに対して所定の条件を満たす第2のテーブルの地理的属性のレコード数を算出してもよい。そして、特徴量生成器82は、算出したレコード数を特徴量として第1のテーブルの属性に追加してもよい。

[0162] このように、特徴量生成器82は、生成した特徴量を第1のテーブルの属性に追加する処理も行うことから、特徴量生成器82のことを属性追加手段とすることができる。また、特徴量生成器82が生成した特徴量は、後述する特徴量選択器83が特徴量を選択する際の候補となることから、特徴量の候補と言うこともできる。

- [0163] なお、本実施形態では、特徴量生成器 8 2 が、特徴量生成関数を用いて特徴量の候補を生成する場合について説明した。ただし、特徴量生成器 8 2 が、類似度関数を用いて、第 1 のテーブルおよび第 2 のテーブルから、結合条件と集約条件とを用いて特徴量の候補を直接生成してもよい。上述するように、結合条件は、第 1 の属性の値と第 2 の属性の値とにより算出される類似度が、条件を満たす第 1 の属性の値を含む第 1 のテーブルのレコードと、第 2 の属性の値を含む第 2 のテーブルのレコードとを結合するための条件である。また、集約条件は、第 2 のテーブルにおける複数のレコードに対する集約方法およびその集約の対象になる列により表される条件である。
- [0164] 特徴量生成器 8 2 は、例えば、結合条件および集約条件がそれぞれ複数存在する場合、複数の結合条件と複数の集約条件とを組み合わせた数の特徴量を生成してもよい。結合条件および集約条件を組み合わせることにより、上述する特徴量生成関数生成器 8 1 が、特徴量生成関数を生成する処理と同様の効果が得られる。
- [0165] 特徴量選択器 8 3 は、生成された特徴量の中から、予測に最適な特徴量を選択する。なお、特徴量選択の方法は任意である。特徴量選択器 8 3 は、例えば、L 1 正則化を用いて特徴量を選択してもよい。ただし、特徴量の選択に用いるアルゴリズムはL 1 正則化に限られない。特徴量選択器 8 3 は、特徴量の選択に用いるアルゴリズムに応じて、予測に最適な特徴量を選択すればよい。
- [0166] 出力部 9 0 は、生成された特徴量を出力する。出力部 9 0 は、特徴量選択器 8 3 が選択した特徴量のみを出力してもよく、特徴量生成器 8 2 が生成した全ての特徴量を出力してもよい。
- [0167] 学習部 9 1 は、生成された特徴量を用いて予測モデルを学習する。学習部 9 1 は、例えば、追加された属性を特徴量として予測モデルを学習する。具体的には、学習部 9 1 は、生成された特徴量に第 1 のテーブルおよび第 2 のテーブルのデータを適用して、訓練データを生成する。そして、学習部 9 1 は、生成された特徴量を説明変数の候補として用いて、予測対象の値を予測

するモデルを学習する。なお、モデルの学習方法は任意である。

[0168] 予測部 9 2 は、学習部 9 1 によって学習されたモデルを用いて予測を行う。具体的には、予測部 9 2 は、生成された特徴量に第 1 のテーブルおよび第 2 のテーブルのデータを適用して、予測用データを生成する。そして、予測部 9 2 は、生成された予測用データを学習されたモデルに適用して予測結果を得る。

[0169] 入力部 1 0 と、ジオコード 2 0 と、マップパラメータ生成器 3 0 と、フィルタパラメータ生成器 5 0 と、集約パラメータ生成器 6 0 と、特徴量生成関数生成器 8 1 と、特徴量生成器 8 2 と、特徴量選択器 8 3 と、出力部 9 0 と、学習部 9 1 と、予測部 9 2 とは、プログラム（情報処理プログラム）に従って動作するコンピュータのプロセッサ（例えば、CPU（Central Processing Unit）、GPU（Graphics Processing Unit）、FPGA（field-programmable gate array））によって実現される。なお、マップパラメータ生成器 3 0 は、より詳しくは、ジオマップ生成器 4 0（さらに詳しくは、距離マップ生成器 4 1 と、包含マップ生成器 4 2 と、重複マップ生成器 4 3 と、同地域マップ生成器 4 4）と、時間差異マップ生成器 3 1 と、マップ生成器 3 2 と、属性特定部 3 3 とにより実現される。また、集約パラメータ生成器 6 0 は、ジオ集約生成器 7 0（さらに詳しくは、ポイント集約生成器 7 1 と、エリア集約生成器 7 2）と、数集約生成器 6 1 とにより実現される。

[0170] 例えば、プログラムは、記憶部 8 0 に記憶され、プロセッサは、そのプログラムを読み込み、プログラムに従って、入力部 1 0、ジオコード 2 0、マップパラメータ生成器 3 0、フィルタパラメータ生成器 5 0、集約パラメータ生成器 6 0、特徴量生成関数生成器 8 1、特徴量生成器 8 2、特徴量選択器 8 3、出力部 9 0、学習部 9 1 および予測部 9 2 として動作してもよい。また、情報処理システムの機能が SaaS（Software as a Service）形式で提供されてもよい。

[0171] 入力部 1 0 と、ジオコード 2 0 と、マップパラメータ生成器 3 0 と、フィルタパラメータ生成器 5 0 と、集約パラメータ生成器 6 0 と、特徴量生成関

数生成器 8 1 と、特徴量生成器 8 2 と、特徴量選択器 8 3 と、出力部 9 0 と、学習部 9 1 と、予測部 9 2 とは、それぞれが専用のハードウェアで実現されていてもよい。また、各装置の各構成要素の一部又は全部は、汎用または専用の回路 (circuitry)、プロセッサ等やこれらの組合せによって実現されてもよい。これらは、単一のチップによって構成されてもよいし、バスを介して接続される複数のチップによって構成されてもよい。各装置の各構成要素の一部又は全部は、上述した回路等とプログラムとの組合せによって実現されてもよい。

[0172] また、各装置の各構成要素の一部又は全部が複数の情報処理装置や回路等により実現される場合には、複数の情報処理装置や回路等は、集中配置されてもよいし、分散配置されてもよい。例えば、情報処理装置や回路等は、クライアントアンドサーバシステム、クラウドコンピューティングシステム等、各々が通信ネットワークを介して接続される形態として実現されてもよい。また、本実施形態の情報処理システム 1 0 0 が、単体の情報処理装置として実現されていてもよい。また、本実施形態の情報処理システム 1 0 0 の一部または全部は、上述する特徴量を生成する処理を行うことから、特徴量を生成する処理を行う機能を含む装置を、特徴量生成装置とすることができる。

[0173] 次に、本実施形態の情報処理システム 1 0 0 の動作を説明する。図 1 9 は、結合条件を生成する処理の例を示すフローチャートである。

[0174] 入力部 1 0 は、予測対象および第 1 の地理的属性を含む第 1 のテーブルと、第 2 の地理的属性を含む第 2 のテーブルとを取得する (ステップ S 1 1)。また、入力部 1 0 は、地理的關係、および、地理的關係の程度を受け付ける (ステップ S 1 2)。マップパラメータ生成器 3 0 は、第 1 の地理的属性の値と第 2 の地理的属性の値との関係が地理的關係の程度を満たすような、第 1 のテーブルに含まれるレコードと第 2 のテーブルに含まれるレコードとを結合するための結合条件を生成する (ステップ S 1 3)。

[0175] 図 2 0 は、結合条件を生成する処理の他の例を示すフローチャートである

。入力部 10 は、予測対象および第 1 の時間的属性を含む第 1 のテーブルと、第 2 の時間的属性を含む第 2 のテーブルとを取得する（ステップ S 2 1）。また、入力部 10 は、時間的關係、および、時間的關係の程度を受け付ける（ステップ S 2 2）。マップパラメータ生成器 30 は、第 1 の時間的属性の値と第 2 の時間的属性の値との關係が時間的關係の程度を満たすような、第 1 のテーブルに含まれるレコードと第 2 のテーブルに含まれるレコードとを結合するための結合条件を生成する（ステップ S 2 3）。

[0176] 図 2 1 は、特徴量を生成する処理の例を示すフローチャートである。入力部 10 は、予測対象および第 1 の地理的属性を含む第 1 のテーブルと、第 2 の地理的属性を含む第 2 のテーブルとを取得する（ステップ S 3 1）。特徴量生成器 8 2 は、第 1 の地理的属性の値に対する第 2 の地理的属性の値が所定の条件を満たす場合に距離の統計値を算出し（ステップ S 3 2）、算出した統計値を特徴量として第 1 のテーブルの属性に追加する（ステップ S 3 3）。

[0177] 図 2 2 は、特徴量を生成する処理の他の例を示すフローチャートである。入力部 10 は、予測対象および第 1 の属性を含む第 1 のテーブルと、第 2 の属性を含む第 2 のテーブルとを取得する（ステップ S 4 1）。また、入力部 10 は、第 1 の属性と第 2 の属性との類似度の算出に用いられる類似度関数と、類似度に対する条件（例えば、類似度の閾値）とを受け付ける（ステップ S 4 2）。特徴量生成器 8 2 は、類似度関数を用いて算出される結合条件と集約条件とを用いて、第 1 のテーブルおよび第 2 のテーブルから特徴量の候補を生成する（ステップ S 4 3）。そして、特徴量選択器 8 3 は、特徴量の候補から、予測に最適な特徴量を選択する（ステップ S 4 4）。

[0178] 以上のように、本実施形態では、入力部 10 が、予測対象および第 1 の地理的属性を含む第 1 のテーブルと、第 2 の地理的属性を含む第 2 のテーブルとを取得する。また、入力部 10 は、地理的關係、および、地理的關係の程度を受け付ける。そして、マップパラメータ生成器 30 が、第 1 の地理的属性の値と第 2 の地理的属性の値との關係が地理的關係の程度を満たすような

、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するための結合条件を生成する。同様に、本実施形態では、入力部10が、予測対象および第1の時間的属性を含む第1のテーブルと、第2の時間的属性を含む第2のテーブルとを取得する。また、入力部10は、時間的關係、および、時間的關係の程度を受け付ける。そして、マップパラメータ生成器30が、第1の時間的属性の値と第2の時間的属性の値との關係が時間的關係の程度を満たすような、第1のテーブルに含まれるレコードと第2のテーブルに含まれるレコードとを結合するための結合条件を生成する。よって、地理的情報または時間的情報を介して複数の情報を関連付ける作業工数を低減できる。その結果、多様な表現で表された情報を処理するコンピュータの負荷を低減することが可能になる。

[0179] また、本実施形態では、入力部10が、予測対象および第1の地理的属性を含む第1のテーブルと、第2の地理的属性を含む第2のテーブルとを取得する。そして、特徴量生成器82は、第1の地理的属性の値に対する第2の地理的属性の値が所定の条件を満たす場合に、第1の地理的属性の値と条件を満たす第2の地理的属性の値とに基づいて算出される距離の統計値を、予測対象に影響を及ぼし得る変数である特徴量として第1のテーブルの属性に追加する。よって、地理的情報を有する複数の情報源から、効率よく特徴量を生成できる。

[0180] さらに、本実施形態では、入力部10が、予測対象および第1の属性を含む第1のテーブルと、第2の属性を含む第2のテーブルとを取得する。また、入力部10が、第1の属性と第2の属性との類似度の算出に用いられる類似度関数と、類似度に対する条件とを受け付ける。そして、特徴量生成器82が、類似度関数を用いて算出される結合条件と集約条件とを用いて、第1のテーブルおよび第2のテーブルから特徴量の候補を生成し、特徴量選択器83が、特徴量の候補から、予測に最適な特徴量を選択する。よって、特徴量を生成するための分析者工数を削減できる。

[0181] 次に、本発明の概要を説明する。図23は、本発明による特徴量生成装置

の概要を示すブロック図である。本発明による特徴量生成装置 380 は、予測対象および第 1 の属性を含む第 1 のテーブル（例えば、ターゲットテーブル）と、第 2 の属性を含む第 2 のテーブル（例えば、ソーステーブル）とを取得するテーブル取得手段 381（例えば、入力部 10）と、第 1 の属性と第 2 の属性との類似度の算出に用いられる類似度関数と、類似度に対する条件とを受け付ける受付手段 382（例えば、入力部 10）と、類似度関数を用いて、第 1 の属性の値と第 2 の属性の値とにより算出される類似度が、条件（類似度に対する条件）を満たす第 1 の属性の値を含む第 1 のテーブルのレコードと、第 2 の属性の値を含む第 2 のテーブルのレコードとを結合するための結合条件（例えば、マップパラメータ）と、第 2 のテーブルにおける複数のレコードに対する集約方法および集約の対象になる列により表される集約条件（例えば、集約パラメータ）とを用いて、第 1 のテーブルおよび第 2 のテーブルから、予測対象に影響を及ぼし得る特徴量の候補を生成する特徴量生成手段 383（例えば、特徴量生成器 82）と、特徴量の候補から、予測に最適な特徴量を選択する特徴量選択手段 384（例えば、特徴量選択器 83）とを備えている。

[0182] そのような構成により、特徴量を生成するための分析者工数を削減できる。

[0183] また、特徴量生成手段 383 は、複数の結合条件と複数の集約条件とを用いて、結合条件と集約条件とを組み合わせた数の特徴量を生成してもよい。

[0184] その際、第 1 の属性および第 2 の属性が、地理的属性であり、類似度関数が、距離が近いほど類似度を高く算出する関数として定義されてもよい。

[0185] 他にも、第 1 の属性および第 2 の属性が、時間的属性であり、類似度関数が、時間の差異が小さいほど類似度を高く算出する関数として定義されてもよい。

[0186] 他にも、第 1 の属性および第 2 の属性が、文字列属性であり、類似度関数が、文字列を示すテキストの一致度が高いほど類似度を高く算出する関数として定義されてもよい。

- [0187] 他にも、第1の属性および第2の属性が、構造的属性であり、類似度関数が、共通するノードまでの距離が近いほど類似度を高く算出する関数として定義されてもよい。
- [0188] また、情報処理装置380は、結合条件と、集約条件と、第2のテーブルから条件を満たすレコードを抽出するための抽出条件（例えば、フィルタパラメータ）とを用いて特徴量生成関数を生成する関数生成手段（例えば、特徴量生成関数生成器81）を備えていてもよい。
- [0189] 図24は、少なくとも1つの実施形態に係るコンピュータの構成を示す概略ブロック図である。コンピュータ1000は、プロセッサ1001、主記憶装置1002、補助記憶装置1003、インタフェース1004を備える。
- [0190] 上述の情報処理システムは、コンピュータ1000に実装される。そして、上述した各処理部の動作は、プログラム（結合条件生成プログラム）の形式で補助記憶装置1003に記憶されている。プロセッサ1001は、プログラムを補助記憶装置1003から読み出して主記憶装置1002に展開し、当該プログラムに従って上記処理を実行する。
- [0191] なお、少なくとも1つの実施形態において、補助記憶装置1003は、一時的でない有形の媒体の一例である。一時的でない有形の媒体の他の例としては、インタフェース1004を介して接続される磁気ディスク、光磁気ディスク、CD-ROM、DVD-ROM、半導体メモリ等が挙げられる。また、このプログラムが通信回線によってコンピュータ1000に配信される場合、配信を受けたコンピュータ1000が当該プログラムを主記憶装置1002に展開し、上記処理を実行しても良い。
- [0192] また、当該プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、当該プログラムは、前述した機能を補助記憶装置1003に既に記憶されている他のプログラムとの組み合わせで実現するもの、いわゆる差分ファイル（差分プログラム）であっても良い。
- [0193] 上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが

、以下には限られない。

- [0194] (付記1) 予測対象および第1の属性を含む第1のテーブルと、第2の属性を含む第2のテーブルとを取得するテーブル取得手段と、前記第1の属性と前記第2の属性との類似度の算出に用いられる類似度関数と、当該類似度に対する条件とを受け付ける受付手段と、前記類似度関数を用いて、前記第1の属性の値と前記第2の属性の値とにより算出される類似度が、前記条件を満たす当該第1の属性の値を含む第1のテーブルのレコードと、当該第2の属性の値を含む第2のテーブルのレコードとを結合するための結合条件と、第2のテーブルにおける複数のレコードに対する集約方法および当該集約の対象になる列により表される集約条件とを用いて、前記第1のテーブルおよび前記第2のテーブルから、前記予測対象に影響を及ぼし得る特徴量の候補を生成する特徴量生成手段と、前記特徴量の候補から、予測に最適な特徴量を選択する特徴量選択手段とを備えたことを特徴とする特徴量生成装置。
- [0195] (付記2) 特徴量生成手段は、複数の結合条件と複数の集約条件とを用いて、当該結合条件と当該集約条件とを組み合わせた数の特徴量を生成する付記1記載の特徴量生成装置。
- [0196] (付記3) 第1の属性および第2の属性は、地理的属性であり、類似度関数は、距離が近いほど類似度を高く算出する関数として定義される付記1または付記2記載の特徴量生成装置。
- [0197] (付記4) 第1の属性および第2の属性は、時間的属性であり、類似度関数は、時間の差異が小さいほど類似度を高く算出する関数として定義される付記1または付記2記載の特徴量生成装置。
- [0198] (付記5) 第1の属性および第2の属性は、文字列属性であり、類似度関数は、文字列を示すテキストの一致度が高いほど類似度を高く算出する関数として定義される付記1または付記2記載の特徴量生成装置。
- [0199] (付記6) 第1の属性および第2の属性は、構造的属性であり、類似度関数は、共通するノードまでの距離が近いほど類似度を高く算出する関数として定義される付記1または付記2記載の特徴量生成装置。

- [0200] (付記 7) 結合条件と、集約条件と、第 2 のテーブルから条件を満たすレコードを抽出するための抽出条件とを用いて特徴量生成関数を生成する関数生成手段を備えた付記 1 から付記 6 のうちのいずれか 1 つに記載の特徴量生成装置。
- [0201] (付記 8) 予測対象および第 1 の属性を含む第 1 のテーブルと、第 2 の属性を含む第 2 のテーブルとを取得し、前記第 1 の属性と前記第 2 の属性との類似度の算出に用いられる類似度関数と、当該類似度に対する条件とを受け付け、前記類似度関数を用いて、前記第 1 の属性の値と前記第 2 の属性の値とにより算出される類似度が、前記条件を満たす当該第 1 の属性の値を含む第 1 のテーブルのレコードと、当該第 2 の属性の値を含む第 2 のテーブルのレコードとを結合するための結合条件と、第 2 のテーブルにおける複数のレコードに対する集約方法および当該集約の対象になる列により表される集約条件とを用いて、前記第 1 のテーブルおよび前記第 2 のテーブルから、前記予測対象に影響を及ぼし得る特徴量の候補を生成し、前記特徴量の候補から、予測に最適な特徴量を選択することを特徴とする特徴量生成方法。
- [0202] (付記 9) 複数の結合条件と複数の集約条件とを用いて、当該結合条件と当該集約条件とを組み合わせた数の特徴量を生成する付記 8 記載の特徴量生成方法。
- [0203] (付記 10) コンピュータに、予測対象および第 1 の属性を含む第 1 のテーブルと、第 2 の属性を含む第 2 のテーブルとを取得するテーブル取得処理、前記第 1 の属性と前記第 2 の属性との類似度の算出に用いられる類似度関数と、当該類似度に対する条件とを受け付ける受付処理、前記類似度関数を用いて、前記第 1 の属性の値と前記第 2 の属性の値とにより算出される類似度が、前記条件を満たす当該第 1 の属性の値を含む第 1 のテーブルのレコードと、当該第 2 の属性の値を含む第 2 のテーブルのレコードとを結合するための結合条件と、第 2 のテーブルにおける複数のレコードに対する集約方法および当該集約の対象になる列により表される集約条件とを用いて、前記第 1 のテーブルおよび前記第 2 のテーブルから、前記予測対象に影響を及ぼし得

る特徴量の候補を生成する特徴量生成処理、および、前記特徴量の候補から、予測に最適な特徴量を選択する特徴量選択処理を実行させるための特徴量生成プログラム。

[0204] (付記 11) コンピュータに、特徴量生成処理で、複数の結合条件と複数の集約条件とを用いて、当該結合条件と当該集約条件とを組み合わせた数の特徴量を生成させる付記 10 記載の特徴量生成プログラム。

[0205] 以上、実施形態及び実施例を参照して本願発明を説明したが、本願発明は上記実施形態および実施例に限定されるものではない。本願発明の構成や詳細には、本願発明のスコープ内で当業者が理解し得る様々な変更をすることができる。

[0206] この出願は、2017年10月5日に出願された米国仮出願第62/568,397号を基礎とする優先権を主張し、その開示の全てをここに取り込む。

符号の説明

- [0207]
- 10 入力部
 - 20 ジオコード
 - 30 マップパラメータ生成器
 - 31 時間差異マップ生成器
 - 32 マップ生成器
 - 33 属性特定部
 - 40 ジオマップ生成器
 - 41 距離マップ生成器
 - 42 包含マップ生成器
 - 43 重複マップ生成器
 - 44 同地域マップ生成器
 - 50 フィルタパラメータ生成器
 - 51 フィルタ生成器
 - 60 集約パラメータ生成器

- 6 1 数的集約生成器
- 7 0 ジオ集約生成器
- 7 1 ポイント集約生成器
- 7 2 エリア集約生成器
- 8 0 記憶部
- 8 1 特徴量生成関数生成器
- 8 2 特徴量生成器
- 8 3 特徴量選択器
- 9 0 出力部
- 9 1 学習部
- 9 2 予測部