

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2017年9月21日(21.09.2017)



(10) 国際公開番号

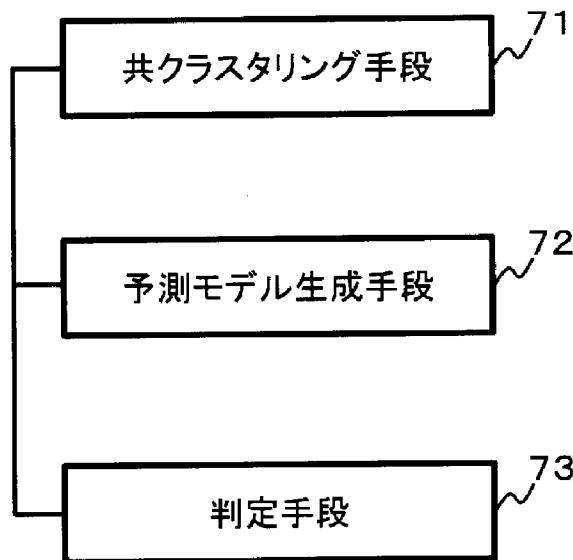
WO 2017/159402 A1

- (51) 国際特許分類:
G06F 17/30 (2006.01) G06N 99/00 (2010.01)
- (21) 国際出願番号: PCT/JP2017/008488
- (22) 国際出願日: 2017年3月3日(03.03.2017)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2016-052737 2016年3月16日(16.03.2016) JP
- (71) 出願人: 日本電気株式会社(NEC CORPORATION)
[JP/JP]; 〒1088001 東京都港区芝五丁目7番1号
Tokyo (JP).
- (72) 発明者: 小山田 昌史(OYAMADA Masafumi); 〒
1088001 東京都港区芝五丁目7番1号 日本電
気株式会社内 Tokyo (JP). 中台 慎二(NAKADAI
Shinji); 〒1088001 東京都港区芝五丁目7番1号
日本電気株式会社内 Tokyo (JP).
- (74) 代理人: 岩壁 冬樹, 外(IWAKABE Fuyuki et al.);
〒1040031 東京都中央区京橋二丁目8番7号
読売八重洲ビル6階 サンライズ国際特許事務
所 Tokyo (JP).
- (81) 指定国 (表示のない限り、全ての種類の国内保
護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA,
BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN,
CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG,
ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL,
IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA,
LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN,
MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE,
PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE,
SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT,
TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 指定国 (表示のない限り、全ての種類の広域保
護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW,
MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユー
ロピア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨー
ロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE,
ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC,

[続葉有]

(54) Title: CO-CLUSTERING SYSTEM, METHOD, AND PROGRAM

(54) 発明の名称: 共クラスタリングシステム、方法およびプログラム



71 Co-clustering means
72 Prediction model generation means
73 Determination means

(57) Abstract: Provided is a co-clustering system with which the prediction accuracy of a prediction model for each cluster can be improved. A co-clustering means 71 executes co-clustering processing to co-cluster a first ID, which is an ID for a record in first master data, and a second ID, which is an ID for a record in second master data, on the basis of the first master data, the second master data, and fact data that indicates the relationship between the first ID and the second ID. A prediction model generation means 72 executes prediction model generation processing to generate a prediction model at least for each cluster of the first ID. A determination means 73 determines whether or not a prescribed condition is satisfied. Prediction model generation processing and co-clustering processing are repeated until the prescribed condition is determined to be satisfied.

(57) 要約: クラスタ毎の予測モデルの予測精度をより向上させることができる共クラスタリングシステムを提供する。共クラスタリング手段71は、第1のマスターデータと、第2のマスターデータと、第1のマスターデータ内のレコードのIDである第1IDと第2のマスターデータ内のレコードのIDである第2IDとの関係を示すファクトデータとに基づいて、第1IDおよび第2IDを共クラスタリングする共クラスタリング処理を実行する。予測モデル生成手段72は、少なくとも第1IDのクラスタ毎に予測モデルを生成する予測モデル生成処理を実行する。判定手段73は、所定の条件が満たされたか否かを判定する。所定の条件が満たされた判定されるまで、予測モデル生成処理および共クラスタリング処理を繰り返す。



MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, 添付公開書類:
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, — 国際調査報告 (条約第 21 条(3))
KM, ML, MR, NE, SN, TD, TG).

明 細 書

発明の名称：共クラスタリングシステム、方法およびプログラム 技術分野

[0001] 本発明は、２種類の事項それぞれをクラスタリングする共クラスタリングシステム、共クラスタリング方法および共クラスタリングプログラムに関する。

背景技術

[0002] 回帰・判別に代表される教師あり学習は、例えば、小売店における商品の需要予測や、電力使用量の予測等、様々な分析処理に用いられる。教師あり学習は、入力と出力との組が与えられると、入力と出力との関係性を学習し、出力の不明な入力を与えられると、学習した関係性に基づいて、その出力を予測する。

[0003] 近年、教師あり学習の予測精度を向上させるために、１つのデータセットに対して複数の予測モデルを生成し、予測時には適切に予測モデルを選択したり、適切にそれらの予測モデルを混合させたりする技術が提案されている。この技術は、Mixture of Expertsと呼ばれる。Mixture of Expertsの１つとして、混合モデルを用いた技術が非特許文献１に記載されている。非特許文献１に記載の技術は、データ（例えば、商品ＩＤ）を、データの性質（例えば、商品の価格）に基づいてクラスタリングし、クラスタ毎に予測モデルを生成する。この結果、同じクラスタに所属する「性質の類似したデータ」に基づいて予測モデルを生成することになる。従って、データ全体で予測モデルを生成する場合と比べ、非特許文献１に記載の技術では、より細部を捉えた予測モデルを生成することができ、予測精度が向上する。

[0004] 以下にこの具体例を示す。

例えば、あるサービスの会員が年間にエステティックサロンを利用する利用回数を、年齢から予測する予測問題を考える。この予測問題は、年齢を入力とし、利用回数を出力する関数を求める問題である。また、ここでは、デ

ータ全体が6人分のデータであるとする。図23は、その6人分の年齢と利用回数とをグラフに示した結果を例示した図である。図23に示すグラフにおいて、 x 軸は年齢を示し、 y 軸は利用回数を示している。また、その6人分のデータ全体から、線形回帰により予測モデル（上記の関数）を生成し、その関数を図示すると、その関数は、図23に示す直線のように示すことができる。この関数に年齢 x を代入したときの y の値が利用回数の予測値となる。図23から分かるように、この予測値と、実際の利用回数との差は大きく、予測精度は低い。

[0005] これに対し、非特許文献1に記載の技術を利用して、6人分のデータを、「美容派」と「酒好き」の2つのクラスタに分けたとする。この場合のクラスタ毎の年齢と利用回数、および予測モデルの例を図24に示す。図24（a）は、「美容派」に対応するグラフであり、図24（b）は、「酒好き」に対応するグラフである。図24においても、 x 軸は年齢を示し、 y 軸は利用回数を示している。図24から分かるように、傾向の同じデータを同じクラスタにまとめて、クラスタ毎に予測モデルを生成することによって、それぞれのクラスタで高い予測精度を実現することができる。

[0006] また、非特許文献2には、IRM（Infinite Relational Model）を用いた学習が記載されている。非特許文献2に記載された学習では、データセット内に未知の値が存在することを許容しない。例えば、学習に用いられるデータセットが、顧客IDと、その顧客の種々の属性の値との組の集合であるとする。非特許文献2に記載された学習では、それらの属性の中に、値が定まっていない属性が存在することを許容しない。

先行技術文献

非特許文献

[0007] 非特許文献1: Jun Zhu, Ning Chen, Eric P. Xing, “Infinite SVM: a Dirichlet Process Mixture of Large-margin Kernel Machines”, In ICML, pages 617-624.

非特許文献2: Charles Kemp, Joshua B. Tenenbaum, Thomas L. Griffiths,

Takeshi Yamada, Naonori Ueda, "Learning Systems of Concepts with an Infinite Relational Model", In AAAI, volume 21, pages 381-338.

発明の概要

発明が解決しようとする課題

[0008] 非特許文献1に記載された技術では、データセット（例えば、顧客情報）をデータ自身が持つ属性の値（例えば、顧客の年齢）を用いてクラスタリングし、属性の似た顧客のクラスタ毎に、未知の属性（例えば、顧客の収入）の予測モデルを生成する。なお、未知の属性は、各データのうち、一部のデータに関して未知であり、この属性の値が分かっているデータも存在しているものとする。上記の例では、顧客の収入が既知となっているデータと、顧客の収入が未知であるデータとが混在しているものとする。そのように予測モデルを生成した結果、各クラスタの特徴をより捉えた予測モデルを生成することができ、予測精度の向上が可能となる。しかし、予測対象となる未知の属性の値と、他の属性の値との相関が小さい場合には、予測精度の向上は望めない。例えば、上記の例において、顧客の年齢と、顧客の年収との間に相関がほとんどない場合、年齢から年収を予測する予測モデルをクラスタ毎に生成したとしても、予測精度の向上は望めない。

[0009] そこで、本発明は、クラスタ毎の予測モデルの予測精度をより向上させることができる共クラスタリングシステム、共クラスタリング方法および共クラスタリング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

[0010] 本発明による共クラスタリングシステムは、第1のマスタデータと、第2のマスタデータと、第1のマスタデータ内のレコードのIDである第1IDと第2のマスタデータ内のレコードのIDである第2IDとの関係を示すファクトデータとに基づいて、第1IDおよび第2IDを共クラスタリングする共クラスタリング処理を実行する共クラスタリング手段と、少なくとも第1IDのクラスタ毎に予測モデルを生成する予測モデル生成処理を実行する予測モデル生成手段と、所定の条件が満たされたか否かを判定する判定手段

とを備え、所定の条件が満たされたと判定されるまで、予測モデル生成処理および共クラスタリング処理を繰り返し、共クラスタリング手段が、1つの第1 IDが1つのクラスタに属する所属確率を決定する際、第1 IDに対応する目的変数の値を、クラスタに対応する予測モデルを用いて予測し、当該値と実際の値との差が小さいほど、所属確率を高い確率とすることを特徴とする。

[0011] また、本発明による共クラスタリング方法は、第1のマスタデータと、第2のマスタデータと、第1のマスタデータ内のレコードのIDである第1 IDと第2のマスタデータ内のレコードのIDである第2 IDとの関係を示すファクトデータとに基づいて、第1 IDおよび第2 IDを共クラスタリングする共クラスタリング処理を実行し、少なくとも第1 IDのクラスタ毎に予測モデルを生成する予測モデル生成処理を実行し、所定の条件が満たされたか否かを判定し、所定の条件が満たされたと判定されるまで、予測モデル生成処理および共クラスタリング処理を繰り返し、共クラスタリング処理で、1つの第1 IDが1つのクラスタに属する所属確率を決定する際、第1 IDに対応する目的変数の値を、クラスタに対応する予測モデルを用いて予測し、当該値と実際の値との差が小さいほど、所属確率を高い確率とすることを特徴とする。

[0012] また、本発明による共クラスタリングプログラムは、コンピュータに、第1のマスタデータと、第2のマスタデータと、第1のマスタデータ内のレコードのIDである第1 IDと第2のマスタデータ内のレコードのIDである第2 IDとの関係を示すファクトデータとに基づいて、第1 IDおよび第2 IDを共クラスタリングする共クラスタリング処理、少なくとも第1 IDのクラスタ毎に予測モデルを生成する予測モデル生成処理、および、所定の条件が満たされたか否かを判定する判定処理を実行させ、所定の条件が満たされたと判定されるまで、予測モデル生成処理および共クラスタリング処理を繰り返させ、共クラスタリング処理で、1つの第1 IDが1つのクラスタに属する所属確率を決定する際、第1 IDに対応する目的変数の値を、クラス

タに対応する予測モデルを用いて予測させ、当該値と実際の値との差が小さいほど、所属確率を高い確率とさせることを特徴とする。

発明の効果

[0013] 本発明によれば、クラスタ毎の予測モデルの予測精度をより向上させることができる。

図面の簡単な説明

[0014] [図1]第1のマスターデータの例を示す説明図である。

[図2]第2のマスターデータの例を示す説明図である。

[図3]ファクトデータの例を示す説明図である。

[図4]ハードクラスタリングの結果の例を示す模式図である。

[図5]ソフトクラスタリングの結果の例を示す模式図である。

[図6]本発明の第1の実施形態の共クラスタリングシステムの例を示す機能ブロック図である。

[図7]予測モデル学習部が学習モデルを生成する際に用いる教師データの説明図である。

[図8]クラスタ関係の例を示す模式図である。

[図9]クラスタ関係の例を示す模式図である。

[図10]ファクトデータの例を示す模式図である。

[図11]第1の実施形態の処理経過の例を示すフローチャートである。

[図12]図1、図2に示す第1のマスターデータ、第2のマスターデータ、および図3に示すファクトデータを統合した結果の例を示す説明図である。

[図13]第1のマスターデータの例を示す説明図である。

[図14]第2のマスターデータの例を示す説明図である。

[図15]ファクトデータの例を示す説明図である。

[図16]本発明の第2の実施形態の予測システムの例を示す機能ブロック図である。

[図17]第2の実施形態の処理経過の例を示すフローチャートである。

[図18]本発明の第3の実施形態の共クラスタリングシステムの例を示す機能

ブロック図である。

[図19]第1の実施形態の具体例における処理経過の例を示すフローチャートである。

[図20]第1の実施形態の具体例における処理経過の例を示すフローチャートである。

[図21]本発明の各実施形態に係るコンピュータの構成例を示す概略ブロック図である。

[図22]本発明の共クラスタリングシステムの概要を示すブロック図である。

[図23]6人分の年齢と利用回数とをグラフに示した結果を例示した図である。

[図24]6人分のデータを2つのクラスタに分け、クラスタ毎に年齢と利用回数とをグラフに示した結果を例示した図である。

発明を実施するための形態

[0015] 以下、本発明の実施形態を図面を参照して説明する。

[0016] まず、本発明において、事前に与えられるデータについて説明する。本発明では、第1のマスタデータ、第2のマスタデータおよびファクトデータが与えられる。なお、マスタデータは、ディメンションデータと称される場合もある。従って、第1のマスタデータ、第2のマスタデータをそれぞれ、第1のディメンションデータ、第2のディメンションデータと称してもよい。また、ファクトデータは、トランザクションデータまたは実績データと称される場合もある。

[0017] 第1のマスタデータおよび第2のマスタデータは、それぞれ複数のレコードを含む。第1のマスタデータのレコードのIDを第1IDと記す。第2のマスタデータのレコードのIDを第2IDと記す。

[0018] 第1のマスタデータの各レコードでは、第1IDと、その第1IDに対応する属性の値とが対応付けられている。ただし、第1IDに対応する属性のうち、特定の属性に関しては、一部のレコードで値が未知である。

[0019] 第2のマスタデータの各レコードでは、第2IDと、その第2IDに対応

する属性の値とが対応付けられている。なお、第2 IDに対応する属性のうち、特定の属性に関しては、一部のレコードで値が未知であってもよい。ただし、以下の説明では、第2のマスタデータでは、各属性の値は全て定められている場合を例にして説明する。

[0020] ここでは、第1 IDが顧客 IDであり、第2 IDが商品 IDである場合を例示して説明する。第1 IDおよび第2 IDは、顧客 IDや商品 IDに限定されるわけではない。

[0021] 図1は、第1のマスタデータの例を示す説明図である。図1では、値が未知であることを“?”で表している。図1では、顧客 ID（第1 ID）に対応する属性として、「年齢」、「年収」、「年間のエステティックサロンの利用回数」を例示している。「顧客1」および「顧客2」のレコードでは、「年間のエステティックサロンの利用回数」の値が定められている。しかし、「顧客3」および「顧客4」のレコードでは、「年間のエステティックサロンの利用回数」の値が未知となっている。一部のレコードで値が未知になる状況は、例えば、一部の顧客からのみ、アンケートで「年間のエステティックサロンの利用回数」の回答を得た場合等に生じる。他の属性（「年齢」、「年収」）の値は、各レコードで定められている。なお、図1に例示したマスタデータは、顧客データであると言える。

[0022] 図2は、第2のマスタデータの例を示す説明図である。図2では、商品 ID（第2 ID）に対応する属性として、「商品名」および「価格」を例示している。図2に示す各属性の値は全て定められている。なお、図2に例示したマスタデータは、商品データであると言える。

[0023] ファクトデータは、第1 IDと第2 IDとの関係を示すデータである。図3は、ファクトデータの例を示す説明図である。図3に示す例では、顧客 ID（第1 ID）によって特定される顧客が、商品 ID（第2 ID）によって特定される商品を購入した実績があるか否かという関係を示している。図3では、顧客が商品を購入した実績があることを“1”で示し、実績がないことを“0”で示している。例えば、図3に示す例では、「顧客1」は、「商

品 1」を購入したことはあるが、「商品 2」を購入したことはない。なお、ファクトデータにおいて、第 1 ID と第 2 ID との関係を示す値は 2 値（“0” および “1”）に限定されない。例えば、顧客 ID と商品 ID との関係を示す値は、顧客が商品を購入した個数等であってもよい。図 3 に例示するファクトデータは、購買実績データであると言える。

[0024] また、本発明の各実施形態の説明の前にクラスタリングについて説明する。クラスタリングとは、データをクラスタと呼ばれる複数のグループに分けるタスクである。クラスタリングでは、データに何等かの性質が定められ、性質が似たデータが同じクラスタに所属するように、データが分けられる。クラスタリングには、ハードクラスタリングと、ソフトクラスタリングとがある。

[0025] ハードクラスタリングでは、個々のデータはいずれか 1 つのクラスタのみに所属させられる。図 4 は、ハードクラスタリングの結果の例を示す模式図である。

[0026] ソフトクラスタリングでは、個々のデータは複数のクラスタに所属させられる。このとき、各データには、クラスタ毎に、「クラスタにどの程度所属しているか」を表す所属確率が割り当てられる。図 5 は、ソフトクラスタリングの結果の例を示す模式図である。

[0027] なお、ハードクラスタリングは、個々のデータの所属確率がそれぞれ 1 つのクラスタで “1. 0” となり、残りの全てのクラスタで “0. 0” となるクラスタリングと捉えることができる。すなわち、ハードクラスタリングの結果も、2 値の所属確率で表すことができる。また、ハードクラスタリングの結果を導出する過程で、0. 0 ～ 1. 0 の範囲の所属確率を用いてもよい。最終的に、そのような所属確率が最大となるクラスタで所属確率を “1. 0” とし、他の各クラスタの所属確率を “0. 0” にする処理を、各データに関してそれぞれ行えばよい。

[0028] 各実施形態では、特に言及のない限り、ハードクラスタリングとソフトクラスタリングを区別せずに説明する。また、ハードクラスタリングでの所属

クラスタの決定や、ソフトクラスタリング（ハードクラスタリングでもよい。）での所属確率の決定を、クラスタ割り当ての決定と記す。

[0029] 実施形態 1.

本発明の発明者は、非特許文献 2 に記載の I R M を利用して、第 1 のマスタデータ、第 2 のマスタデータおよびファクトデータが与えられた場合に、第 1 I D および第 2 I D を共クラスタリングする処理を検討した。以下、この処理の流れを述べ、さらに、本発明の第 1 の実施形態において、第 1 のマスタデータ、第 2 のマスタデータおよびファクトデータが与えられた場合に、第 1 I D および第 2 I D を共クラスタリングする処理について述べる。

[0030] 第 1 I D および第 2 I D の共クラスタリングでは、第 1 I D の各クラスタと、第 2 I D の各クラスタとの間に（クラスタの直積空間上に）、確率モデルを保持する。確率モデルは、典型的には、クラスタ間の関係の強さを表すベルヌーイ分布である。一方の I D（例えば、第 1 I D）のあるクラスタへの所属確率を算出する際には、そのクラスタと、他方の I D（本例では、第 2 I D）の各クラスタとの間にある確率モデルの値を参照する。例えば、確率モデルとしてクラスタ間の関係の強さを利用した場合、ある顧客 I D がある顧客 I D クラスタに所属する確率は、その顧客 I D クラスタと関係の強い商品 I D クラスタに所属する商品 I D が示す商品を、その顧客 I D が示す顧客がどれだけ購入しているかによって定まる。このように共クラスタリングを実行することによって、似た商品を買う顧客の顧客 I D が同じ顧客 I D クラスタに集まり、また、似た顧客によって買われる商品の商品 I D が同じ商品 I D クラスタに集まる。

[0031] [非特許文献 2 に記載の I R M を利用した共クラスタリング処理]

非特許文献 2 に記載の I R M を利用した共クラスタリング処理では、以下のステップを繰り返す。

[0032] 1. 第 1 I D の各クラスタ（第 1 I D を要素とする各クラスタ）への所属確率、および、第 2 I D の各クラスタ（第 2 I D を要素とする各クラスタ）への所属確率を更新する。所属確率は、ファクトデータ（例えば、図 3 に例示

する購買実績データ）と、第１ＩＤや第２ＩＤに対応する属性（例えば、顧客の年齢や商品の価格）とから定まる。

[0033] ２．

（２－１）第１ＩＤの各クラスタの重み（事前確率）、および、第２ＩＤの各クラスタの重み（事前確率）を更新する。例えば、第１のマスターデータ（図１参照）の中に若い人のレコードが多い場合、若年層のクラスタに第１ＩＤが所属する事前確率を高くする。

（２－２）第１ＩＤを要素とする各クラスタ、および第２ＩＤを要素とする各クラスタを対象にして、クラスタのモデル情報を、現時点でのクラスタ割り当てに基づいて更新する。クラスタのモデル情報は、そのクラスタに所属するＩＤに対応する属性の値の統計的な性質を表す情報である。クラスタのモデル情報は、そのクラスタの代表的な要素の持つ性質を表現していると言える。例えば、クラスタのモデル情報は、クラスタに所属しているＩＤに対応する属性の値の平均や分散で表すことができる。なお、第１ＩＤの各クラスタへの所属確率および第２ＩＤの各クラスタへの所属確率が判明しているので、クラスタのモデル情報（例えば、顧客の平均年齢や商品の平均価格）を計算することができる。

[0034] ３．第１ＩＤの各クラスタと、第２ＩＤの各クラスタとの間に保持される確率モデルを、各ＩＤの所属確率に基づいて更新する。例えば、ある顧客ＩＤクラスタと、ある商品ＩＤクラスタとの関係は、それらのクラスタに属する顧客ＩＤと商品ＩＤとの間の関係（例えば、購買実績）が存在するほど、強くなる。

[0035] 上記の“１．”～“３．”の各ステップを繰り返し、その繰り返しが必要なくなったと判定した時点で、共クラスタリングの処理を終了する。

[0036] [本発明の第１の実施形態の共クラスタリング処理]

本発明の第１の実施形態の共クラスタリング処理では、一部のレコードで特定の属性の値が未知となっているマスターデータ（ここでは、第１のマスターデータ）における各レコードのＩＤ（すなわち、第１ＩＤ）のクラスタ毎に

、予測モデルを保持する。本実施形態では、属性の値が類似している第1 IDを同じクラスタに所属させ、クラスタ毎に異なる予測モデルを生成することで、上記の特定の属性における未知の値の予測精度を向上させる。また、本実施形態では、クラスタ割り当ての決定において、第1 IDが各クラスタに所属する所属確率を、クラスタに対応する予測モデルの予測誤差が小さいほど高い確率とすることで、クラスタリングの精度を向上させる。

[0037] 本発明の第1の実施形態の共クラスタリング処理では、以下のステップを繰り返す。

[0038] 1. 第1 IDの各クラスタで、クラスタに所属する第1 IDに対応する属性の値を用いて、予測モデルを更新する。例えば、サポートベクタマシンの重みを更新する。

[0039] 2. 第1 IDの各クラスタ（第1 IDを要素とする各クラスタ）への所属確率、および、第2 IDの各クラスタ（第2 IDを要素とする各クラスタ）への所属確率を更新する。所属確率は、ファクトデータ（例えば、図3に例示する購買実績データ）と、第1 IDや第2 IDに対応する属性（例えば、顧客の年齢や商品の価格）とから定まる。第1 IDの各クラスタへの所属確率を定める場合には、クラスタ毎の予測モデルも考慮される。例えば、ある第1 IDに関して、予測モデルによる予測精度が高いクラスタほど、その第1 IDの所属確率を高くする。

[0040] 3.

（3-1）第1 IDの各クラスタの重み（事前確率）、および、第2 IDの各クラスタの重み（事前確率）を更新する。例えば、第1のマスタデータ（図1参照）の中に若い人のレコードが多い場合、若年層のクラスタに第1 IDが所属する事前確率を高くする。

（3-2）第1 IDを要素とする各クラスタ、および第2 IDを要素とする各クラスタを対象にして、クラスタのモデル情報を、現時点でのクラスタ割り当てに基づいて更新する。なお、第1 IDの各クラスタへの所属確率および第2 IDの各クラスタへの所属確率が判明しているので、クラスタのモデ

ル情報（例えば、顧客の平均年齢や商品の平均価格）を計算することができる。

[0041] 4. 第1 IDの各クラスと、第2 IDの各クラスとの間に保持される確率モデルを、各 IDの所属確率に基づいて更新する。例えば、ある顧客 ID クラスと、ある商品 ID クラスとの関係は、それらのクラスに属する顧客 ID と商品 ID との関係（例えば、購買実績）が存在するほど、強くなる。

[0042] 上記の“1.”～“4.”の各ステップを繰り返し、その繰り返しが必要なくなったと判定した時点で、共クラスタリングの処理を終了する。

[0043] 以下、本発明の第1の実施形態について、より具体的に説明する。図6は、本発明の第1の実施形態の共クラスタリングシステムの例を示す機能ブロック図である。

[0044] 本発明の第1の実施形態の共クラスタリングシステム1は、データ入力部2と、処理部3と、記憶部4と、結果出力部5とを備える。処理部3は、初期化部31と、クラスタリング部32とを備える。クラスタリング部32は、予測モデル学習部321と、クラスタ割り当て部322と、クラスタ情報算出部323と、クラスタ関係算出部324と、終了判定部325とを備える。

[0045] データ入力部2は、共クラスタリングに用いられるデータ群と、クラスタリングの設定値とを取得する。データ入力部2は、例えば、外部の装置にアクセスして、データ群と、クラスタリングの設定値とを取得してもよい。あるいは、データ入力部2は、データ群と、クラスタリングの設定値とが入力される入力インタフェースであってもよい。

[0046] 共クラスタリングに用いられるデータ群は、第1のマスタデータ（例えば、図1に例示する顧客データ）と、第2のマスタデータ（例えば、図2に例示する商品データ）と、ファクトデータ（例えば、図3に例示する購買実績データ）とを含む。第1のマスタデータの属性のうち、特定の属性に関しては、一部のレコードで値が未知である。なお、非特許文献2に記載された技

術では、入力されるデータにおいて、値が定まっていない属性が存在することを許容しない。すなわち、非特許文献2に記載された技術では、属性の欠損値を許容しない。従って、一部のレコードで特定の属性の値が未知であるという点は、非特許文献2に記載された技術と異なる。

[0047] クラスタリングの設定値は、例えば、第1IDのクラスタ数の最大値、第2IDのクラスタ数の最大値、予測モデルを生成するマスタデータの指定、予測モデルで説明変数とする属性、予測モデルで目的変数とする属性、および予測モデルの種類を含む。

[0048] 予測モデルは、値が定まっていない特定の属性の値を予測するために用いられる。従って、本例では、予測モデルを生成するマスタデータとして、第1のマスタデータが指定される。予測モデルで目的変数とする属性として、その特定の属性（例えば、図1に示す「年間のエステティックサロンの利用回数」）が指定される。

[0049] 予測モデル種類には、例えば、サポートベクタマシン、サポートベクタ回帰、ロジスティック回帰等がある。予測モデルの種類として、各種予測モデルのうち、いずれかが指定される。

[0050] 初期化部31は、データ入力部2から、第1のマスタデータと、第2のマスタデータと、ファクトデータと、クラスタリングの設定値を受け取り、それらを記憶部4に記憶させる。また、初期化部31は、クラスタリングに用いる各種パラメータを初期化する。

[0051] クラスタリング部32は、繰り返し処理により、第1IDおよび第2IDの共クラスタリングを実現する。以下、クラスタリング部32が備える各部について説明する。なお、予測モデルを生成するマスタデータとして、第1のマスタデータが指定されているものとする。

[0052] 予測モデル学習部321は、予測モデルを生成するマスタデータ（第1のマスタデータ）に関するクラスタ毎に（すなわち、第1IDのクラスタ毎に）、目的変数に該当する属性の予測モデルを学習する。

[0053] クラスタリングがハードクラスタリングである場合、予測モデル学習部3

21は、クラスタに対応する予測モデルを生成するときに、そのクラスタに所属する第1IDに対応する属性の値を教師データとして利用する。

[0054] 図7は、予測モデル学習部321が学習モデルを生成する際に用いる教師データの説明図である。例えば、ハードクラスタリングにより、図7に示す顧客1, 2がクラスタ1のみに所属し、図7に示す顧客3がクラスタ2のみに所属することになったとする。この場合、予測モデル学習部321は、顧客1, 2に対応する各属性の値を教師データとして、クラスタ1に対応する予測モデルを生成し、顧客3に対応する各属性の値を教師データとして、クラスタ2に対応する予測モデルを生成する。

[0055] また、クラスタリングがソフトクラスタリングである場合、予測モデル学習部321は、クラスタに対応する予測モデルを生成するときに、未知の値を含まない全てのレコードの属性の値を教師データとして利用する。このとき、予測モデル学習部321は、各レコードの属性の値を、各第1IDのそのクラスタへの所属確率によって重み付けし、重み付けした結果を用いて、予測モデルを生成する。従って、そのクラスタへの所属確率が高い第1IDに対応する教師データは、そのクラスタに対応する予測モデル内で強く影響し、そのクラスタへの所属確率が低い第1IDに対応する教師データは、その予測モデル内であまり影響しない。

[0056] 図7を用いて具体例を説明する。ソフトクラスタリングでは、図7に示す顧客1, 2, 3はそれぞれの所属確率でクラスタ1に所属する。また、図7に示す顧客1, 2, 3はそれぞれの所属確率でクラスタ2にも所属する。予測モデル学習部321は、クラスタ1に対応する予測モデルを生成する場合、顧客1, 2, 3の各属性の値を、顧客1, 2, 3それぞれのクラスタ1への所属確率で重み付けし、重み付けした結果を用いて予測モデルを生成する。クラスタ2に対応する予測モデルを生成する場合も同様である。

[0057] クラスタ割り当て部322は、それぞれの第1IDおよびそれぞれの第2IDに対して、クラスタ割り当てを行う。クラスタ割り当て部322は、第1IDおよび第2IDを共クラスタリングしているということもできる。な

お、既に説明したように、ハードクラスタリングの結果も、2 値の所属確率で表すことができる。また、ハードクラスタリングの結果を導出する過程で、0. 0～1. 0 の範囲の所属確率を用いてもよい。ここでは、ハードクラスタリングとソフトクラスタリングを区別せずに、所属確率を用いて、クラスタ割り当て部 3 2 2 の動作を説明する。

[0058] クラスタ割り当て部 3 2 2 は、クラスタ割り当てを実行する際、2 つの情報参照する。

[0059] 1 つ目の情報は、ファクトデータである。説明を分かりやすくするために、第 1 ID が顧客 ID であり、第 2 ID が商品 ID である場合を例にして説明する。ある顧客 ID がある顧客 ID クラスタに所属する確率は、その顧客 ID クラスタと関係の強い商品 ID クラスタに所属する商品 ID によって特定される商品を、その顧客 ID によって特定される顧客がどれだけ購入しているかによって定まる。ある商品 ID がある商品 ID クラスタに所属する確率に関しても同様である。クラスタ割り当て部 3 2 2 は、第 1 ID の各クラスタへの所属確率や、第 2 ID の各クラスタへの所属確率を求めるときに、ファクトデータを参照する。この動作の詳細については、後述する。

[0060] また、2 つ目の情報は、予測モデルの精度である。顧客 ID クラスタ（第 1 ID のクラスタ）毎に予測モデルが生成されている。クラスタ割り当て部 3 2 2 は、顧客 ID クラスタに所属する顧客 ID に対応するレコードを、その顧客 ID クラスタに対応する予測モデルに適用して、目的変数となる属性の予測値を計算し、その予測値と正解値（レコードに示されている実際の値）との差を計算する。この差が、予測モデルの精度である。クラスタ割り当て部 3 2 2 は、この差が小さいほど、着目している顧客 ID クラスタに所属している顧客 ID の所属確率を高め、この差が大きいほど、着目している顧客 ID クラスタに所属している顧客 ID の所属確率を低くするように、顧客 ID の所属確率を補正する。クラスタ割り当て部 3 2 2 は、この補正を、各顧客 ID クラスタに対して行う。この動作によって、予測モデルの精度がよくなるように、クラスタリング結果が調節される。

- [0061] クラスタ情報算出部323は、各第1IDおよび各第2IDのクラスタ割り当て（所属確率）を参照し、第1IDの各クラスタおよび第2IDの各クラスタのモデル情報を算出し、記憶部4に記憶されている各クラスタのモデル情報を更新する。既に説明したように、クラスタのモデル情報は、そのクラスタに所属するIDに対応する属性の値の統計的な性質を表す情報である。例えば、各顧客IDクラスタにおいて、各顧客の年収が正規分布に従うとした場合、各顧客IDクラスタのモデル情報は、正規分布における平均値および分散値となる。
- [0062] クラスタのモデル情報は、クラスタ割り当ての決定と、後述するクラスタ関係の計算に利用される。
- [0063] クラスタ関係算出部324は、第1IDの各クラスタと、第2IDの各クラスタとの間のクラスタ関係を算出し、記憶部4に記憶されているクラスタ関係を更新する。クラスタ関係とは、クラスタの組み合わせの性質を表す値である。以下、クラスタ関係が0～1の範囲の値である場合を例にして説明する。クラスタ関係算出部324は、ファクトデータを基に、第1IDのクラスタと第2IDのクラスタの組み合わせ毎に、クラスタ関係を算出する。従って、第1のクラスタの数と第2IDのクラスタの数との積だけ、クラスタ関係が算出される。図8は、クラスタ関係の例を示す模式図である。図8に示す例では、顧客IDクラスタの数が2であり、商品IDクラスタの数が2であるので、クラスタ関係の数は、 $2 * 2 = 4$ となっている。なお、図8に示す「美容好き」、「美容商品」等は、クラスタの内容に基づいて、システム管理者が便宜的に付加したラベルであるものとする。
- [0064] 第1IDのクラスタに所属している第1IDと、第2IDのクラスタに所属している第2IDとの関係性が強い程、その2つのクラスタの組み合わせにおけるクラスタ関係は大きな値となる。例えば、顧客IDクラスタに所属している顧客IDによって特定される顧客と、商品IDクラスタに所属している商品IDによって特定される商品との関係性が強い程、クラスタ関係は“1”に近づき、その関係性が弱いほど、クラスタ関係は“0”に近づく。

図8に示す例において、顧客IDクラスタ1には、美容好きの顧客の顧客IDが多く所属している。また、顧客IDクラスタ2には、酒好きの顧客の顧客IDが多く所属している。また、商品IDクラスタ1には、美容商品の商品IDが多く所属している。例えば、顧客IDクラスタ1と商品IDクラスタ1との間のクラスタ関係は0.9であり、1に近い値である。このことは、顧客IDクラスタ1に所属している顧客IDによって特定される顧客が、商品IDクラスタ1に所属している商品IDによって特定される商品を購入することが多いということ（関係性が強いこと）を表している。また、顧客IDクラスタ2と商品IDクラスタ1との間のクラスタ関係は0.1であり、0に近い値である。このことは、顧客IDクラスタ2に所属している顧客IDによって特定される顧客が、商品IDクラスタ1に所属している商品IDによって特定される商品を購入することが少ないということ（関係性が弱いこと）を表している。

[0065] クラスタ関係算出部324は、以下に示す式(A)を計算することによって、クラスタ関係を算出すればよい。

[0066] [数1]

$$\frac{\hat{a}_{k_1 k_2}^{[1]}}{\hat{a}_{k_1 k_2}^{[1]} + \hat{b}_{k_1 k_2}^{[1]}}$$

(A)

[0067] 式(A)において、 k_1 は、第1IDのクラスタのIDを表し、 k_2 は、第2IDのクラスタのIDを表す。また、 $a_{k_1 k_2}^{[1]}$ 、 $b_{k_1 k_2}^{[1]}$ は、クラスタ関係の計算に用いるパラメータである。 $a_{k_1 k_2}^{[1]}$ が大きいほど、 k_1 と k_2 の関係は強く、 $b_{k_1 k_2}^{[1]}$ が大きいほど k_1 と k_2 の関係は弱い。なお、本明細書の文章内では、数式内で示したハット記号を省略する。

[0068] クラスタ関係算出部324は、 $a_{k_1 k_2}^{[1]}$ を、以下に示す式(B)によっ

て計算すればよい。また、クラスタ関係算出部324は、 $b^{[1]}_{k_1 k_2}$ を、以下に示す式(C)によって計算すればよい。

[0069] [数2]

$$\hat{a}^{[1]}_{k_1 k_2} = a_0^{[1]} + \sum_{d_1}^{D^{(1)}} \sum_{d_2}^{D^{(2)}} \phi_{d_1, k_1}^{(1)} \phi_{d_2, k_2}^{(2)} x_{d_1 d_2} \quad (B)$$

[0070] [数3]

$$\hat{b}^{[1]}_{k_1 k_2} = b_0^{[1]} + \sum_{d_1}^{D^{(1)}} \sum_{d_2}^{D^{(2)}} \phi_{d_1, k_1}^{(1)} \phi_{d_2, k_2}^{(2)} (1 - x_{d_1 d_2}) \quad (C)$$

[0071] 式(B)および式(C)では、 d_1 が、第1IDの順番を表し、 $D^{(1)}$ が、第1IDの総数を表しているものとする。同様に、 d_2 が、第2IDの順番を表し、 $D^{(2)}$ が、第2IDの総数を表しているものとする。式(B)および式(C)において、 $\phi_{d_1, k_1}^{(1)}$ は、 d_1 番目の第1IDがクラスタ k_1 に所属している確率である。 $\phi_{d_2, k_2}^{(2)}$ は、 d_2 番目の第2IDがクラスタ k_2 に所属している確率である。 $x_{d_1 d_2}$ は、 d_1 と d_2 との組み合わせに応じたファクトデータ内の値である。

[0072] ここで、前述のクラスタ割り当て部322がファクトデータを参照して、IDのクラスタへの所属確率を求める処理について、詳細に説明する。ここでは、顧客ID(第1ID)を変数*i*で表す。また、商品ID(第2ID)を変数*j*で表す。また、顧客IDクラスタのIDを変数 k_1 で表す。商品IDクラスタのIDを変数 k_2 で表す。

[0073] また、図9に例示するクラスタ関係が得られているとする。 $k_1 = 1$ である

クラスタは、甘党の顧客の顧客IDを多く含んでいるものとする。 $k_1 = 2$ であるクラスタは、辛党の顧客の顧客IDを多く含んでいるものとする。 $k_2 = 1$ であるクラスタは、甘い商品の商品IDを多く含んでいるものとする。 $k_2 = 2$ であるクラスタは、辛い商品の商品IDを多く含んでいるものとする。 $k_2 = 3$ であるクラスタは、苦い商品の商品IDを多く含んでいるものとする。また、図9に示す「甘党」、「甘い」等は、クラスタの内容に基づいて、システム管理者が便宜的に付加したラベルであるものとする。

[0074] また、図10に例示するファクトデータが与えられているとする。

[0075] ここでは、 $i = 1$ である顧客が、 $k_1 = 2$ であるクラスタに所属する確率をクラスタ割り当て部322が算出する場合を例にして説明する。また、 i がクラスタ k_1 に所属する確率を $q(z_i^{(1)} = k_1)$ と記す。よって、 $i = 1$ である顧客が、 $k_1 = 2$ であるクラスタに所属する確率は、 $q(z_1^{(1)} = 2)$ と表される。また、 j がクラスタ k_2 に所属する確率を $q(z_j^{(2)} = k_2)$ と記す。

[0076] クラスタ割り当て部322は、以下に示す式(D)の計算により、 $q(z_1^{(1)} = 2)$ を求める。

[0077] [数4]

$$q(z_1^{(1)} = 2) \propto \sum_j \sum_{k_2} \left(z_j^{(2)} = k_2 \right) E_q \left[\log p(x_{i=1, j}) \mid \theta_{k_1=2, k_2} \right] \quad (D)$$

[0078] 式(D)において、 x は、添え字の i, j の組み合わせに対応するファクトデータ(図10参照)内の値である。従って、図10に示す例では、 x は、1または0である。また、 θ は、添え字の k_1, k_2 の組み合わせに対応するクラスタ関係である。

[0079] E_q は、確率の期待値を求める演算であり、 $E_q[\log p(x_{i=1, j}) \mid \theta_{k_1=2, k_2}]$ は、 j がクラスタ k_2 に所属しているものとした場合に、顧客 $i = 1$ がその商品 j を買う確率の期待値である。

- [0080] クラスタ割り当て部322は、同様の計算により、着目している顧客IDが、他の顧客IDクラスタに所属している確率も求める。ハードクラスタリングの場合、クラスタ割り当て部322は、その結果得られた所属確率が最高となっている顧客IDクラスタのみに、着目している顧客IDが所属していると決定すればよい。クラスタ割り当て部322は、他の顧客IDに関しても、各顧客IDクラスタに所属している確率を計算する。
- [0081] また、クラスタ割り当て部322は、それぞれの商品IDがそれぞれの商品IDクラスタに所属している確率も、同様の計算によって求める。
- [0082] また、上記の所属確率の算出の後に、クラスタ割り当て部322は、予測モデルを用いた所属確率の補正を実行すればよい。
- [0083] クラスタリング部32は、予測モデル学習部321による処理、クラスタ割り当て部322による処理、クラスタ情報算出部323による処理、および、クラスタ関係算出部324による処理を繰り返す。
- [0084] 終了判定部325は、上記の一連の処理の繰り返しを終了するか否かを判定する。終了判定部325は、終了条件が満たされた場合に、上記の一連の処理の繰り返しを終了すると判定し、終了条件が満たされていなければ、繰り返しを続けると判定する。以下、終了条件の例を説明する。
- [0085] 例えば、上記の一連の処理の繰り返し回数が、クラスタリングの設定値の中で定められていてもよい。終了判定部325は、上記の一連の処理の繰り返し回数が定められた回数に達したときに、繰り返しを終了すると判定してもよい。
- [0086] また、例えば、クラスタ割り当て部322が、クラスタ割り当ての決定を実行するときに、クラスタリングの精度を導出し、そのクラスタリングの精度を記憶部4に記憶させてもよい。終了判定部325は、前回に導出されたクラスタリングの精度から、直近に導出されたクラスタリングの精度への変化量を計算し、その変化量が小さければ（具体的には、変化量の絶対値が所定の閾値以下であれば）、繰り返しを終了すると判定してもよい。
- [0087] なお、ソフトクラスタリングの場合、クラスタ割り当て部322は、クラ

スタリングの精度として、例えば、クラスタリングのモデルの尤度を計算してもよい。また、ハードクラスタリングの場合、クラスタ割り当て部 3 2 2 は、クラスタリングの精度として、例えば、Pseudo F を計算してもよい。

[0088] 記憶部 4 は、データ入力部 2 が取得した種々のデータや、処理部 3 の処理で得られた種々のデータを記憶する記憶装置である。記憶部 4 は、計算機の主記憶装置であっても、二次記憶装置であってもよい。記憶部 4 が二次記憶装置である場合には、クラスタリング部 3 2 は、処理を途中で中断し、その後、再開することができる。また、記憶部 4 が、主記憶装置と二次記憶装置とに分かれた構成であって、処理部 3 は、データの一部を主記憶装置に記憶させ、他のデータを二次記憶装置に記憶させてもよい。

[0089] 結果出力部 5 は、記憶部 4 に記憶された、クラスタリング部 3 2 による処理の結果を出力する。具体的には、結果出力部 5 は、処理の結果として、予測モデル、クラスタ割り当て、クラスタ関係、クラスタのモデル情報の全部または一部を出力する。クラスタ割り当ては、個々の第 1 ID の各クラスタへの所属確率および個々の第 2 ID の各クラスタへの所属確率である。また、ハードクラスタリングの場合、クラスタ割り当ては、個々の第 1 ID がどのクラスタに所属しているかを直接示す情報、および個々の第 2 ID がどのクラスタに所属しているかを直接示す情報であってもよい。

[0090] また、結果出力部 5 が結果を出力する態様は、特に限定されない。例えば、結果出力部 5 は、結果を他の装置に出力してもよい。また、例えば、結果出力部 5 は、結果をディスプレイ装置に表示させてもよい。

[0091] 予測モデル学習部 3 2 1、クラスタ割り当て部 3 2 2、クラスタ情報算出部 3 2 3、クラスタ関係算出部 3 2 4 および終了判定部 3 2 5 を含むクラスタリング部 3 2、並びに、データ入力部 2、初期化部 3 1、結果出力部 5 は、例えば、プログラム（共クラスタリングプログラム）に従って動作するコンピュータの CPU によって実現される。この場合、CPU は、例えば、コンピュータのプログラム記憶装置（図 6 において図示略）等のプログラム記録媒体からプログラムを読み込み、そのプログラムに従って、データ入力部

2、初期化部31、クラスタリング部32、および結果出力部5として、動作すればよい。

[0092] また、図6に示した共クラスタリングシステム1内の各要素が、それぞれ専用のハードウェアで実現されていてもよい。

[0093] また、本発明のシステム1は、2つ以上の物理的に分離した装置が有線または無線で接続されている構成であってもよい。この点は、後述の各実施形態においても同様である。

[0094] 次に、第1の実施形態の処理経過を説明する。図11は、第1の実施形態の処理経過の例を示すフローチャートである。

[0095] データ入力部2は、共クラスタリングに用いられるデータ群（第1のマスタデータ、第2のマスタデータおよびファクトデータ）と、クラスタリングの設定値とを取得する（ステップS1）。

[0096] 初期化部31は、第1のマスタデータ、第2のマスタデータおよびファクトデータと、クラスタリングの設定値を記憶部4に記憶させる。また、初期化部31は、「クラスタのモデル情報」、「クラスタ割り当て」および「クラスタ関係」に対して初期値を設定し、その初期値を記憶部4に記憶させる（ステップS2）。

[0097] ステップS2における初期値は任意であってもよい。あるいは、初期化部31は、各初期値を、例えば、以下に示すように導出してもよい。

[0098] 初期化部31は、第1のマスタデータにおける属性の値の平均値を算出し、その平均値を、第1IDの全てのクラスタにおけるクラスタのモデル情報として定めてもよい。同様に、初期化部31は、第2のマスタデータにおける属性の値の平均値を算出し、その平均値を、第2IDの全てのクラスタにおけるクラスタのモデル情報として定めてもよい。

[0099] 初期化部31は、クラスタ割り当ての初期値を以下のように定めてもよい。ハードクラスタリングの場合、初期化部31は、各第1IDをいずれかのクラスタにランダムに割り当て、同様に、各第2IDもいずれかのクラスタにランダムに割り当てる。また、ソフトクラスタリングの場合、初期化部3

1 は、個々の第 1 ID に対して、各クラスタへの所属確率を均一に定める。
例えば、第 1 ID のクラスタの数が 2 つである場合、各第 1 ID の 1 番目のクラスタへの所属確率および 2 番目の所属確率をそれぞれ 0.5 に定める。
同様に、初期化部 31 は、個々の第 2 ID に対して、各クラスタへの所属確率を均一に定める。

[0100] 初期化部 31 は、第 1 ID のクラスタと第 2 ID のクラスタの組み合わせ毎に、クラスタ関係を同じ値（例えば、0.5 等）に定めてもよい。

[0101] ステップ S2 の後、クラスタリング部 32 は、終了条件が満たされるまで、ステップ S3～S7 の処理を繰り返す。以下、ステップ S3～S7 の処理を説明する。

[0102] 予測モデル学習部 321 は、記憶部 4 に記憶されている情報を参照し、第 1 ID のクラスタ毎に、第 1 のマスタデータ内の一部のレコードで値が未知となっている属性を目的変数とする予測モデルを学習する。そして、予測モデル学習部 321 は、学習によって得た各予測モデルを記憶部 4 に記憶させる（ステップ S3）。

[0103] クラスタ割り当て部 322 は、記憶部 4 に記憶されている各第 1 ID のクラスタ割り当ておよび第 2 ID のクラスタ割り当てを更新する（ステップ S4）。ステップ S4 において、クラスタ割り当て部 322 は、記憶部 4 に記憶されているクラスタ割り当て、ファクトデータ、およびクラスタ関係を読み込み、それらに基づいて、各第 1 ID のクラスタ割り当ておよび第 2 ID のクラスタ割り当てを新たに定める。

[0104] また、予測モデルが生成されている各クラスタに関しては、クラスタ割り当て部 322 は、クラスタに対応する予測モデルを用いて目的変数となる属性の予測値を計算し、その予測値と正解値との差（予測モデルの精度）を計算する。クラスタ割り当て部 322 は、この差が小さいほど、着目しているクラスタに所属している第 1 ID の所属確率を高め、この差が大きいほど、着目しているクラスタに所属している第 1 ID の所属確率を低くするように、第 1 ID の所属確率を補正する。クラスタ割り当て部 322 は、予測モデル

ルが生成されていない各クラスタ（すなわち、第2 IDの各クラスタ）に対しては、この処理を行わなくてよい。

[0105] クラスタ割り当て部322は、更新後の各第1 IDのクラスタ割り当ておよび各第2 IDのクラスタ割り当てを記憶部4に記憶させる。

[0106] 次に、クラスタ情報算出部323は、第1のマスタデータ、および各第1 IDのクラスタの割り当てを参照し、第1 IDのクラスタ毎に、クラスタに属する第1 IDに対応する属性の値を用いて、クラスタのモデル情報を計算し直す。同様に、クラスタ情報算出部323は、第2のマスタデータ、および各第2 IDのクラスタ割り当てを参照し、第2 IDのクラスタ毎に、クラスタに属する第2 IDに対応する属性の値を用いて、クラスタのモデル情報を計算し直す。クラスタ情報算出部323は、記憶部4に記憶されているクラスタのモデル情報を、新たに計算したクラスタのモデル情報で更新する（ステップS5）。

[0107] 次に、クラスタ関係算出部324は、各第1 IDのクラスタ割り当ておよび各第2 IDのクラスタ割り当て、並びにファクトデータを参照し、第1 IDのクラスタと第2 IDのクラスタの組み合わせ毎に、クラスタ関係を計算し直す。クラスタ関係算出部324は、記憶部4に記憶されているクラスタ関係を、新たに計算したクラスタ関係で更新する（ステップS6）。

[0108] 次に、終了判定部325は、終了条件が満たされたか否かを判定する（ステップS7）。終了条件が満たされていない場合（ステップS7のNo）、終了判定部325は、ステップS3～S7を繰り返すと判定する。そして、クラスタリング部32は、ステップS3～S7を再度、実行する。

[0109] また、終了条件が満たされた場合（ステップS7のYes）、終了判定部325は、ステップS3～S7の繰り返しを終了すると判定する。この場合、結果出力部5は、その時点におけるクラスタリング部32による処理の結果を出力し、共クラスタリングシステム1の処理が終了する。

[0110] 本実施形態によれば、クラスタ割り当て部322は、ファクトデータを参照して、第1 IDおよび第2 IDのクラスタ割り当てを行う。換言すれば、

クラスタ割り当て部322は、ファクトデータを参照して、第1IDおよび第2IDの共クラスタリングを実行する。そして、予測モデル学習部321は、クラスタ毎に予測モデルを生成する。この結果、クラスタ毎に異なる予測モデルが得られる。また、ファクトデータは、第1IDと第2IDとの関係を表している。例えば、ファクトデータは、「顧客1」は「商品1」を購入したことがあるが、「商品2」は購入したとことがない等の関係を表している。従って、本実施形態における第1IDのクラスタリング結果は、単に第1のマスタデータ内の属性の値に基づいて第1IDをクラスタリングした場合のクラスタリング結果と比較して、より適切なクラスタが得られる。第2IDのクラスタリング結果に関しても同様である。そのような、より適切なクラスタ毎に個別に予測モデルが得られるので、クラスタ毎の予測モデルの予測精度をより向上させることができる。

[0111] また、本実施形態では、予測モデル学習部321は、クラスタの予測精度に応じて、クラスタに所属するIDの所属確率を調節する。このことから、より適切なクラスタが得られる。よって、クラスタ毎の予測モデルの予測精度をより向上させることができる。

[0112] また、上記の説明では、図1に例示する顧客データにおいて、一部のレコードで特定の属性の値が未知である場合を例にして説明した。顧客データ内では、各属性の値が全て定まっていて、図2に例示する商品データにおいて、一部のレコードで特定の属性の値が未知となってもよい。この場合、商品データを第1のマスタデータとし、顧客データを第2のマスタデータとして、共クラスタリングシステム1は、第1の実施形態と同様の処理を行えばよい。

[0113] また、第1のマスタデータと第2のマスタデータそれぞれにおいて、一部のレコードで特定の属性の値が未知となってもよい。この場合、予測モデル学習部321は、第1IDのクラスタ毎に予測モデルを学習し、第2IDのクラスタ毎に予測モデルを学習すればよい。また、クラスタ割り当て部322は、第2IDに関しても、各クラスタへの所属確率を定める際に、第

２ＩＤのクラスタに対応する予測モデルの精度を用いればよい。

[0114] また、第１のマスターデータ、第２のマスターデータおよびファクトデータに基づいて予測モデルを生成する方法として、上述の第１の実施形態による方法とは別に、以下の方法が考えられる。具体的には、第１のマスターデータの各レコードに、第２のマスターデータおよびファクトデータが示す情報を追加することで、第１のマスターデータ、第２のマスターデータおよびファクトデータを統合し、クラスタリングは行わずに、統合後のデータに基づいて予測モデルを学習する方法が考えられる。しかし、この方法で得られる予測モデルの予測精度は、上述の第１の実施形態で得られる予測モデルの予測精度よりも低い。この点について、具体的に説明する。

[0115] 図１２は、図１、図２に示す第１のマスターデータ、第２のマスターデータ、および図３に示すファクトデータを統合した結果の例を示す説明図である。

「炭酸水」、「焼酎」等の商品名に該当する列には、ファクトデータ（図３参照）に基づいて“１”または“０”が格納される。“１”は顧客が商品を購入したことがあることを意味し、“０”は顧客が商品を購入したことがないことを意味する。また、図１２では、「炭酸水」、「焼酎」等の商品名の隣の列には、その商品の価格が格納される場合を例示している。

[0116] 図１２に示す統合結果は、顧客ＩＤ以外の各列は顧客ＩＤの属性となる形式で表されている。このことは、統合前のマスターデータが示していた一部の情報が失われることを意味する。例えば、図１２に示す例では、炭酸水の価格は、本来、顧客ＩＤの属性ではないが、形式的に、顧客ＩＤの属性として表される。そして、炭酸水の価格が顧客ＩＤの属性として扱われるため、「炭酸水」の価格が「１５０」とあるという、統合前の第２のマスターデータ（図２参照）で示されていた情報が、失われることになる。

[0117] 従って、図１２に示す統合結果に基づいて予測モデルを生成したとしても、その予測モデルの予測精度は、上述の第１の実施形態で得られる予測モデルの予測精度よりも低い。

[0118] 実施形態２．

本発明の第2の実施形態では、共クラスタリングを実行し、第1 IDのクラスタ毎に予測モデルを生成し、さらに、予測モデルによる予測を実行する予測システムについて説明する。

[0119] 本発明の第2の実施形態の予測システムにも、第1のマスタデータ、第2のマスタデータおよびファクトデータが入力される。第2の実施形態における第1のマスタデータ、第2のマスタデータおよびファクトデータはそれぞれ、第1の実施形態における第1のマスタデータ、第2のマスタデータおよびファクトデータと同様である。

[0120] 第1のマスタデータにおいて、第1 IDに対応する属性のうち、特定の属性に関しては、一部のレコードで値が未知である。

[0121] また、第2の実施形態では、第2のマスタデータでは、各属性の値は全て定められているものとする。

[0122] また、第2の実施形態では、第1 ID（第1のマスタデータのレコードのID）が顧客IDであり、第1のマスタデータは、顧客と、その顧客の属性との対応関係を表しているものとする。また、第2 ID（第2のマスタデータのレコードのID）が商品IDであり、第2のマスタデータは、商品と、その商品の属性との対応関係を表しているものとする。

[0123] なお、顧客IDは、顧客を表しているので、顧客IDを単に顧客と称してもよい。同様に、商品IDは、商品を表しているので、商品IDを単に商品と称してもよい。

[0124] 以下、第2の実施形態では、図13に例示する第1のマスタデータ、および図14に例示する第2のマスタデータを参照して説明する。第1のマスタデータでは、図13に示す属性以外の属性が示されていてもよい。第2のマスタデータでは、図14に示す属性以外の属性が示されていてもよい。

[0125] ファクトデータは、第1 ID（顧客ID）と第2 ID（商品ID）との関係を示すデータである。第2の実施形態では、ファクトデータは、顧客が商品を購入した実績があるか否かという関係を示しているものとする。図3に示す場合と同様に、顧客が商品を購入した実績があることを“1”で示し、

実績がないことを“0”で示すものとする。

[0126] 以下、第2の実施形態では、図15に例示するファクトデータを参照して説明する。

[0127] 図16は、本発明の第2の実施形態の予測システムの例を示す機能ブロック図である。本発明の第2の実施形態の予測システム500は、共クラスタリング部501と、予測モデル生成部502と、予測部503とを備える。

[0128] 予測システム500には、第1のマスタデータ、第2のマスタデータおよびファクトデータが入力される。

[0129] 共クラスタリング部501は、第1のマスタデータ、第2のマスタデータおよびファクトデータに基づいて、第1ID（顧客ID）および第2ID（商品ID）を共クラスタリングする。共クラスタリング部501は、第1のマスタデータ、第2のマスタデータおよびファクトデータに基づいて、顧客および商品を共クラスタリングするということもできる。

[0130] 共クラスタリング部501が第1のマスタデータ、第2のマスタデータおよびファクトデータに基づいて顧客IDおよび商品IDを共クラスタリングする方法は、公知の共クラスタリング方法でよい。また、共クラスタリング部501は、共クラスタリングとして、ソフトクラスタリングを実行しても、ハードクラスタリングを実行してもよい。

[0131] 第1の実施形態では、所定の条件が満たされたと判定されるまで、予測モデルの生成と、共クラスタリング処理を繰り返す（より具体的には、ステップS3～S7の処理を繰り返す）処理を示したが、第2の実施形態では、そのような繰り返しを行わない場合を例にして説明する。従って、第2の実施形態では、後述の予測モデル生成部502は、共クラスタリング部501による顧客IDおよび商品IDの共クラスタリングの完了後に、予測モデルの生成を行う。

[0132] 予測モデル生成部502は、共クラスタリング部501による共クラスタリングが完了すると、顧客IDのクラス毎に、予測モデルを生成する。

[0133] このとき、予測モデル生成部502は、一部のレコードで値が未知となっ

ている第1のマスタデータ内の属性を目的変数とする予測モデルを生成する。例えば、予測モデル生成部502は、図13に示す「年間のエステティックサロンの利用回数」を目的変数とする予測モデルを生成する。

[0134] また、予測モデル生成部502は、未知の値がない第1のマスタデータ内の属性の一部または全部を説明変数とする予測モデルを生成する。例えば、予測モデル生成部502は、図13に示す「年齢」や「年収」等を説明変数とする予測モデルを生成する。予測モデル生成部502は、例えば、「年齢」のみ（あるいは、「年収」のみ）を説明変数とする予測モデルを生成してもよい。

[0135] さらに、予測モデル生成部502は、第1のマスタデータ内の属性だけでなく、第2のマスタデータ内の属性の値から算出される集約値を説明変数として用いてもよい。ただし、予測モデル生成部502は、第2のマスタデータ内の属性の値から算出される集約値を説明変数として用いる場合、ファクトデータによって顧客IDとの関連があると判定される第2のマスタデータ内の各レコードにおける属性の値の統計量を説明変数とする。

[0136] 「ファクトデータによって顧客IDとの関連があると判定される第2のマスタデータ内の各レコードにおける属性の値の統計量」の例として、例えば、「顧客が購入した商品の価格のうちの最大値」、「顧客が購入した商品の価格の平均値」等が挙げられるが、これらに限定されない。上記の例において、「顧客が購入した商品」は、ファクトデータによって顧客IDとの関連があると判定される第2のマスタデータ内のレコードに該当する。予測モデル生成部502は、そのようなレコードにおける価格の統計量（例えば、最大値、平均値等）を説明変数として用いてもよい。以下、「顧客が購入した商品の価格のうちの最大値」を説明変数として用いる場合を例にして説明する。

[0137] 予測モデル生成部502は、説明変数の値および目的変数の値を特定可能な顧客IDに着目して、説明変数の値および目的変数の値を特定し、それらの値を教師データとして用いて機械学習を実行することによって、予測モデ

ルを生成すればよい。予測モデル生成部502は、この処理をクラス毎に行えばよい。

[0138] 例えば、図13に示す「顧客3」に対応する目的変数（年間のエステティックサロンの利用回数）の値は未知であるので、「顧客3」のレコードは教師データとして用いられない。

[0139] 一方、図13に示す「顧客1」や「顧客2」に関しては、説明変数および目的変数を特定可能である。例えば、「顧客1」や「顧客2」の「年齢」、「年収」等の値、および「年間のエステティックサロンの利用回数」は、第1のマスタデータから特定可能である。さらに、ファクトデータ（図15参照）によって、予測モデル生成部502は、「顧客1」が購入した商品が「炭酸飲料P」のみであると判定し、第2のマスタテーブルの「炭酸飲料P」のレコードにおける属性の統計量として、“130”を特定することができる。すなわち、予測モデル生成部502は、ファクトデータを参照することによって、顧客1が購入した商品の価格のうちの最大値を特定することができる。同様に、ファクトデータ（図15参照）によって、予測モデル生成部502は、「顧客2」が購入した商品が「菓子1」および「炭酸飲料P」であると判定し、第2のマスタテーブルの「菓子1」のレコードおよび「炭酸飲料P」のレコードにおける属性の統計量として、“130”を特定することができる。すなわち、予測モデル生成部502は、ファクトデータを参照することによって、顧客2が購入した商品の価格のうちの最大値を特定することができる。従って、「顧客1」や「顧客2」に関するデータは、教師データとして用いることができる。

[0140] なお、共クラスタリング部501がソフトクラスタリングを実行した場合、教師データの値を、顧客IDが各クラスに所属する所属確率に応じて重み付けすればよい。

[0141] 予測部503は、顧客IDと、目的変数（実施形態では「年間のエステティックサロンの利用回数」という属性）の指定を、例えば、予測システム500のユーザから受け付ける。すると、予測部503は、指定された顧客ID

Dに対応する目的変数の値を、予測モデル生成手段502が生成した予測モデルを用いて予測する。

[0142] 共クラスタリング部501がハードクラスタリングを実行した場合、予測部503は、指定された顧客IDが属するクラスタを特定し、そのクラスタに対応する予測モデルを用いて、その顧客IDに対応する目的変数の値を予測する。

[0143] このとき、予測部503は、指定された顧客IDに対する説明変数の値を特定し、その説明変数の値を、指定された顧客IDが属するクラスタに対応する予測モデルに適用することによって、予測値を算出すればよい。例えば、説明変数が「年齢」および「顧客が購入した商品の価格のうちの最大値」であるとする。また、図13に示す「顧客4」が指定されたとする。予測部503は、第1のマスタデータから「顧客4」の年齢“50”を特定する。また、予測部503は、ファクトデータ（図15参照）によって、「顧客4」が購入した商品が「菓子1」、「炭酸飲料P」および「炭酸飲料Q」であると判定し、「菓子1」、「炭酸飲料P」および「炭酸飲料Q」の価格の最大値“130”を第2のマスタデータ（図14参照）から求める。そして、予測部503は、各説明変数の値“50”，“130”を、「顧客4」が所属するクラスタに対応する予測モデルに適用すればよい。

[0144] また、共クラスタリング部501がソフトクラスタリングを実行した場合、予測部503は、顧客IDの個々のクラスタに対応する予測モデル毎に、指定された顧客IDに対応する目的変数の値を予測する。1つの予測モデルに着目して目的変数の値を予測する動作は上記の動作と同様であり、説明を省略する。

[0145] 予測部503は、個々のクラスタに対応する予測モデル毎に予測値を得た後に、その各予測値を、指定された顧客IDが各クラスタに属する所属確率で重み付け加算し、その結果を目的変数の値として確定する。

[0146] 共クラスタリング部501、予測モデル生成部502および予測部503は、例えば、プログラム（予測プログラム）に従って動作するコンピュータ

のCPUによって実現される。この場合、CPUは、例えば、コンピュータのプログラム記憶装置（図16において図示略）等の等のプログラム記録媒体からプログラムを読み込み、そのプログラムに従って、共クラスタリング部501、予測モデル生成部502および予測部503として動作すればよい。また、共クラスタリング部501、予測モデル生成部502および予測部503がそれぞれ専用のハードウェアで実現されていてもよい。

[0147] 次に、第2の実施形態の処理経過を説明する。図17は、第2の実施形態の処理経過の例を示すフローチャートである。

[0148] 予測システム500に第1のマスタデータ、第2のマスタデータおよびファクトデータが入力されると、共クラスタリング部501は、第1のマスタデータ、第2のマスタデータおよびファクトデータに基づいて、顧客IDおよび商品IDを共クラスタリングする（ステップS101）。ステップS101における共クラスタリングの方法は、公知の共クラスタリング方法でよい。共クラスタリング部501は、共クラスタリングの結果得た各クラスタを予測モデル生成部502に出力する。

[0149] 顧客IDおよび商品IDの共クラスタリングが完了すると、予測モデル生成部502は、共クラスタリング部501が出力した顧客IDのクラスタ毎に、予測モデルを生成する（ステップS102）。予測モデル生成部502の動作の詳細については、既に説明したので、ここでは説明を省略する。

[0150] ステップS102の後、予測部503は、顧客IDと、目的変数の指定を受け付けると、指定された顧客IDに対応する目的変数の値を、ステップS102で生成された予測モデルを用いて予測する（ステップS103）。予測部503の動作の詳細については、既に説明したので、ここでは説明を省略する。

[0151] 第2の実施形態によれば、共クラスタリング部501は、第1のマスタデータ、第2のマスタデータおよびファクトデータに基づいて、顧客ID（第1ID）および商品ID（第2ID）を共クラスタリングする。従って、顧客ID、商品IDそれぞれのクラスタリング精度は、第1のマスタデータだ

けに基づいて顧客IDをクラスタリングする場合や、第2のマスタデータに基づいて商品IDをクラスタリングする場合に比べ向上する。

[0152] そのような良好な精度でクラスタリングされた顧客IDのクラスタ毎に、予測モデル生成部502は、予測モデルを生成する。従って、予測モデルの精度も良好となり、予測モデルに基づいて得られた目的変数の予測値の精度も高くなる。すなわち、第2の実施形態の予測システムによれば、高い精度で予測を行うことができる。

[0153] また、予測モデル生成部502は、第1のマスタデータの属性だけでなく、ファクトデータによって顧客IDとの関連があると判定される第2のマスタデータ内の各レコードにおける属性の値の統計量も、予測モデルの説明変数として用いることが好ましい。そのような統計量も説明変数として用いることで、予測モデルの精度をさらに向上させることができ、その結果、予測モデルに基づいて得られた予測値の精度もさらに向上する。

[0154] 実施形態3.

第2の実施形態では、第1の実施形態とは異なり、予測モデルの生成と、共クラスタリング処理との繰り返しをせずに、共クラスタリングが完了した後に予測モデルを生成するシステムを説明した。

[0155] 本発明の第3の実施形態の共クラスタリングシステムは、第1の実施形態と同様に、ステップS3～S7の処理を繰り返すことにより、第1IDおよび第2IDを共クラスタリングするとともに、クラスタに対応する予測モデルを生成する。さらに、本発明の第3の実施形態の共クラスタリングシステムは、テストデータが入力されると、目的変数の値を予測する。

[0156] 図18は、本発明の第3の実施形態の共クラスタリングシステムの例を示す機能ブロック図である。第1の実施形態と同様の要素については、図6と同一の符号を付し、説明を省略する。第3の実施形態の共クラスタリングシステム1は、データ入力部2、処理部3、記憶部4および結果出力部5に加え、さらに、テストデータ入力部6と、予測部7と、予測結果出力部8とを備える。

- [0157] 以下の説明では、処理部3が、第1の実施形態で説明した処理を完了し、第1 IDおよび第2 IDがそれぞれクラスタに分類され、第1 IDのクラスタ毎に予測モデルが生成されているものとして説明する。
- [0158] テストデータ入力部6は、テストデータを取得する。テストデータ入力部6は、例えば、外部の装置にアクセスして、テストデータを取得してもよい。あるいは、テストデータ入力部6は、テストデータが入力される入力インタフェースであってもよい。
- [0159] テストデータは、目的変数（例えば、図1に示す第1のマスタデータにおける「年間のエステティックサロンの利用回数」）が未知である新たな第1 IDのレコードと、その新たな第1 IDと第2のマスタデータ内の第2 IDとの関係を示すデータとを含む。
- [0160] 新たな第1 IDのレコードは、例えば、あるサービスに会員登録して間もない会員のレコードである。このレコードにおいて、目的変数に該当する属性以外の属性（例えば、「年齢」、「年収」等）の値は定められているものとする。
- [0161] また、その新たな第1 IDと第2のマスタデータ内の第2 IDとの関係を示すデータの例として、その新たな第1 IDによって特定される顧客の商品購買履歴データが挙げられる。新たな第1 IDと第2のマスタデータ内の第2 IDとの関係を示すデータは、新たな第1 IDに関するファクトデータであるということもできる。
- [0162] 予測部7は、テストデータに含まれる新たな第1 IDが所属するクラスタを特定する。このとき、予測部7は、新たな第1 IDのレコードに含まれる属性の値に基づいて、クラスタを特定してもよい。例えば、予測部7は、新たな第1 IDのレコードに含まれる属性の値（例えば、「年齢」、「年収」の値）と、各クラスタに所属する各第1 IDのレコードにおけるその属性の値とを比較して、所属している各第1 IDの属性の値が、新たな第1 IDのレコードに含まれる属性の値に最も近いクラスタを特定してもよい。予測部7は、そのクラスタを、新たな第1 IDが所属するクラスタとみなしてよい。

。

[0163] また、予測部 7 は、新たな第 1 ID と第 2 のマスタデータ内の第 2 ID との関係を示すデータ（例えば、商品購買履歴データ）に基づいて、その新たな第 1 ID によって特定される顧客の商品購買傾向を特定し、同様の商品購買傾向を有する第 1 ID のクラスタを特定してもよい。予測部 7 は、そのクラスタを、新たな第 1 ID が所属するクラスタとみなしてもよい。

[0164] 予測部 7 は、第 1 ID が所属するクラスタを特定した後、そのクラスタに対応する予測モデルに、新たな第 1 ID のレコードに含まれる属性の値を適用することによって、新たな第 1 ID に対応する目的変数の値を予測する。

[0165] 上記の説明では、予測部 7 が、新たな第 1 ID が所属するクラスタを特定する場合を例にして説明した。予測部 7 は、第 1 ID のクラスタ毎に、新たな第 1 ID がクラスタに所属する所属確率を求めてもよい。例えば、予測部 7 は、新たな第 1 ID のレコードに含まれる属性の値（例えば、「年齢」、「年収」の値）と、各クラスタに所属する各第 1 ID のレコードにおけるその属性の値とを比較して、クラスタ毎に、クラスタに所属している各第 1 ID の属性の値と、新たな第 1 ID のレコードに含まれる属性の値との近さの程度に応じて、新たな第 1 ID の各クラスタへの所属確率を求めてもよい。

[0166] また、予測部 7 は、新たな第 1 ID と第 2 のマスタデータ内の第 2 ID との関係を示すデータ（例えば、商品購買履歴データ）に基づいて、その新たな第 1 ID によって特定される顧客の商品購買傾向を特定し、その商品購買傾向と、第 1 ID のクラスタ毎の商品購買傾向との近さの程度に応じて、新たな第 1 ID の各クラスタへの所属確率を求めてもよい。

[0167] 新たな第 1 ID の各クラスタへの所属確率を求めた場合には、予測部 7 は、第 1 ID の各クラスタに対応する予測モデル毎に、新たな第 1 ID のレコードに含まれる属性の値を適用し、目的変数の値を予測する。さらに、予測部 7 は、個々のクラスタに対応する予測モデル毎に予測値を得た後に、その各予測値を、新たな第 1 ID の各クラスタへの所属確率で重み付け加算し、その結果を目的変数の値として確定してもよい。

[0168] 予測結果出力部 8 は、予測部 7 が予測した目的変数の値を出力する。予測結果出力部 8 が目的変数の予測値を出力する態様は、特に限定されない。例えば、予測結果出力部 8 は、目的変数の予測値を他の装置に出力してもよい。また、例えば、予測結果出力部 8 は、目的変数の予測値をディスプレイ装置に表示させてもよい。

[0169] テストデータ入力部 6、予測部 7 および予測結果出力部 8 も、例えば、プログラム（共クラスタリングプログラム）に従って動作するコンピュータの CPU によって実現される。

[0170] 本実施形態によれば、与えられたテストデータにおける未知の値を予測することができる。

[0171] [具体例]

以下に、第 1 の実施形態の具体例を示す。以下に示す具体例では、マスターデータをデータセットと記す場合がある。また、第 1 のマスターデータを“データセット 1”と記し、第 2 のマスターデータを“データセット 2”と記す場合がある。また、ファクトデータを関係データと記す場合がある。

[0172] 以下に示す具体例で示す数式で用いる記号等の意味を、以下に示す表にまとめる。

[0173] [表 1]

記号	説明
ψ	ディガンマ関数
$D^{(1)}$	データセット 1 内のデータ総数
$D^{(2)}$	データセット 2 内のデータ総数
d_1	データセット 1 の d_1 番目のデータを指し示すインデックス
d_2	データセット 2 の d_2 番目のデータを指し示すインデックス
$K^{(1)}$	データセット 1 のクラスタ総数
$K^{(2)}$	データセット 2 のクラスタ総数
k_1	データセット 1 の k_1 番目のクラスタを指し示すインデックス
k_2	データセット 2 の k_2 番目のクラスタを指し示すインデックス
$q(z_{d_1}^{(1)} = k_1) = \phi_{d_1, k_1}^{(1)}$	データセット 1 のデータ d_1 がクラスタ k_1 に所属する確率
$q(z_{d_2}^{(2)} = k_2) = \phi_{d_2, k_2}^{(2)}$	データセット 2 のデータ d_2 がクラスタ k_2 に所属する確率

[0174]

[表2]

$\hat{\alpha}_{k_1}^{(1)}, \hat{\beta}_{k_1}^{(1)}$	データセット 1 のクラスタ k_1 への所属確率の計算につかわれるパラメータ (クラスタ k_1 が選ばれる確率は $\hat{\alpha}_{k_1}^{(1)}$ が大きいほど高く、 $\hat{\beta}_{k_1}^{(1)}$ が大きいほど低く調節される)
$\hat{\alpha}_{k_2}^{(2)}, \hat{\beta}_{k_2}^{(2)}$	データセット 2 のクラスタ k_2 への所属確率の計算につかわれるパラメータ (クラスタ k_2 が選ばれる確率は $\hat{\alpha}_{k_2}^{(2)}$ が大きいほど高く、 $\hat{\beta}_{k_2}^{(2)}$ が大きいほど低く調節される)
$\alpha^{(1)}$	データセット 1 のクラスタ数をコントロールするハイパーパラメータ。この値が大きいほどクラスタ数が大きくなりやすい
$\alpha^{(2)}$	データセット 2 のクラスタ数をコントロールするハイパーパラメータ。この値が大きいほどクラスタ数が大きくなりやすい
$x_{d_1 d_2}$	データセット 1 のデータ d_1 と、2 番目のデータセットの d_2 番目のデータに関係があるかどうかを示す関係データ ($x_{d_1 d_2} = 1$ なら関係あり、0 なら関係なし)
$\hat{a}_{k_1 k_2}^{[1]}, \hat{b}_{k_1 k_2}^{[1]}$	データセット 1 のクラスタ k_1 とデータセット 2 のクラスタ k_2 の間の関係の強さの計算に使われるパラメータ ($\hat{a}_{k_1 k_2}^{[1]}$ が大きいほど関係は強く、 $\hat{b}_{k_1 k_2}^{[1]}$ が大きいほど関係は小さくなる)
$a_0^{[1]}, b_0^{[1]}$	$\hat{a}_{k_1 k_2}^{[1]}, \hat{b}_{k_1 k_2}^{[1]}$ の計算につかわれるハイパーパラメータ。 $a_0^{[1]} > b_0^{[1]}$ であると、クラスタ間の関係が強くなりやすい
ρ	0 から 1 の間の値であり、この値が 0 に近いほどクラスタ所属確率 $\phi_{d_1, k_1}^{(1)}$ の決定における教師あり学習の効果が強くなる
y_1	データセット 1 の予測対象属性の値 (例: “男”)
$Y^{(1)}$	データセット 1 の予測対象属性の値の種類数 (例: 属性値が “男” か “女” のどちらかなら 2)
$\omega_{y_1, d_1}^{(1)}$	データセット 1 のデータ d_1 が予測モデルの学習に与える影響の強さ
$\mathbb{E}_{q(\eta_{k_1}^{(1)})} [\eta_{k_1}^{(1)}]^\top$	データセット 1 の属性予測モデル (SVM の分離平面)
$f_{d_1}^\Delta(y_1)$	予測スコア算出のための入力生成関数
$\sum_{y_1}^{Y^{(1)}} \omega_{y_1, d_1}^{(1)} \mathbb{E}_{q(\eta_{k_1}^{(1)})} [\eta_{k_1}^{(1)}]^\top f_{d_1}^\Delta(y_1)$	データセット 1 のデータ d_1 の属性をクラスタ k_1 の SVM モデルで予測した際の予測スコア (大きいほど予測誤差が小さい)

[0175] 以下に示す具体例では、無限混合ベイズモデルを用いた場合の変分ベイズ法による推論アルゴリズムを記載する。また、第 1 の実施形態等で例示した場合と同様に、第 1 のマスタデータ（データセット 1）が、顧客に関するマスタデータであり、第 2 のマスタデータ（データセット 2）が、商品に関するマスタデータであるものとする。また、第 1 のマスタデータに、一部のレコードで値が未知となっている属性が存在しているものとする。

[0176] d_1 番目の顧客（顧客 ID）がクラスタ k_1 に所属する確率は、以下に示す式（1）で表される。

[0177]

[数5]

$$\begin{aligned}
\log q(z_{d_1}^{(1)} = k_1) &= \log \phi_{d_1, k_1}^{(1)} \\
&\propto \psi(\hat{\alpha}_{k_1}^{(1)}) - \psi(\hat{\alpha}_{k_1}^{(1)} + \hat{\beta}_{k_1}^{(1)}) + \sum_{k'_1=k_1+1}^{K^{(1)}} \left\{ \psi(\hat{\beta}_{k'_1}^{(1)}) - \psi(\hat{\alpha}_{k'_1}^{(1)} + \hat{\beta}_{k'_1}^{(1)}) \right\} \\
&\quad + \rho \sum_{d_2}^{D^{(2)}} \sum_{k_2}^{K^{(2)}} \phi_{d_2, k_2}^{(2)} \left\{ \psi(\hat{a}_{k_1 k_2}^{[1]} + \hat{b}_{k_1 k_2}^{[1]}) + x_{d_1 d_2} \psi(\hat{a}_{k_1 k_2}^{[1]}) + (1 - x_{d_1 d_2}) \psi(\hat{a}_{k_1 k_2}^{[1]}) \right\} \\
&\quad + (1 - \rho) \left\{ \sum_{y_1}^{Y^{(1)}} \omega_{y_1, d_1}^{(1)} \mathbb{E}_{q(\eta_{k_1}^{(1)})} [\eta_{k_1}^{(1)}]^\top f_{d_1}^\Delta(y_1) \right\} \quad (1)
\end{aligned}$$

[0178] d_2 番目の商品（商品 I D）がクラス k_2 に所属する確率は、以下に示す式（2）で表される。

[0179] [数6]

$$\begin{aligned}
\log q(z_{d_2}^{(2)} = k_2) &= \log \phi_{d_2, k_2}^{(2)} \\
&\propto \psi(\hat{\alpha}_{k_2}^{(2)}) - \psi(\hat{\alpha}_{k_2}^{(2)} + \hat{\beta}_{k_2}^{(2)}) + \sum_{k'_2=k_2+1}^{K^{(2)}} \left\{ \psi(\hat{\beta}_{k'_2}^{(2)}) - \psi(\hat{\alpha}_{k'_2}^{(2)} + \hat{\beta}_{k'_2}^{(2)}) \right\} \\
&\quad + \sum_{d_1}^{D^{(1)}} \sum_{k_1}^{K^{(1)}} \phi_{d_1, k_1}^{(1)} \left\{ \psi(\hat{a}_{k_1 k_2}^{[1]} + \hat{b}_{k_1 k_2}^{[1]}) + x_{d_1 d_2} \psi(\hat{a}_{k_1 k_2}^{[1]}) + (1 - x_{d_1 d_2}) \psi(\hat{a}_{k_1 k_2}^{[1]}) \right\} \quad (2)
\end{aligned}$$

[0180] なお、 Ψ は、ディガンマ関数である。 ρ は、システム管理者が設定可能なパラメータであり、 ρ には、0～1の範囲内の値が設定される。 ρ の値が0に近づくほど、共クラスタリングにおける学習の効果が強くなる。すなわち、予測モデルの精度が向上するように I D のクラス k_2 への所属確率が決定されやすくなる。

[0181] 式（1）内の以下の部分は、 d_1 番目の顧客の属性の値を、クラス k_1 の予測モデルで予測した際のスコアを表す。予測誤差が小さいほど、このスコアは大きくなる。すなわち、予測誤差が小さいほど、 d_1 番目の顧客がクラス k_1 に所属する確率は高くなる。

[0182]

[数7]

$$\sum_{y_1}^{Y^{(1)}} \omega_{y_1, d_1}^{(1)} \mathbb{E}_{q(\eta_{k_1}^{(1)})} [\eta_{k_1}^{(1)}]^\top f_{d_1}^\Delta(y_1)$$

[0183] また、データセット1の隠れ変数の生成モデルは、以下に示す式(3)で表される。

[0184] [数8]

$$z_{d_1}^{(1)} \sim \text{Mult}(\phi_{d_1}^{(1)})$$

(3)

[0185] また、そのパラメータの変分事後分布は、以下に示す式(4)で表される。

[0186] [数9]

$$q(v_{k_1}^{(1)}) = \text{Beta}(\hat{\alpha}_{k_1}^{(1)}, \hat{\beta}_{k_1}^{(1)})$$

(4)

[0187] また、そのパラメータの更新式は、以下に示す式(5)、式(6)で表される。

[0188]

[数10]

$$\hat{\alpha}_{k_1}^{(1)} = 1 + \sum_{d_1}^{D^{(1)}} \phi_{d_1, k_1}^{(1)} \quad (5)$$

[0189] [数11]

$$\hat{\beta}_{k_1}^{(1)} = \alpha^{(1)} + \sum_{d_1}^{D^{(1)}} \sum_{k'_1=k_1+1}^{K^{(1)}} \phi_{d_1, k'_1}^{(1)} \quad (6)$$

[0190] また、データセット2に関するパラメータの更新式は、以下に示す式（7）、式（8）で表される。

[0191] [数12]

$$\hat{\alpha}_{k_2}^{(2)} = 1 + \sum_{d_2}^{D^{(2)}} \phi_{d_2, k_2}^{(2)} \quad (7)$$

[0192]

[数13]

$$\hat{\beta}_{k_2}^{(2)} = \alpha^{(2)} + \sum_{d_2}^{D^{(2)}} \sum_{k'_2=k_2+1}^{K^{(2)}} \phi_{d_2, k'_2}^{(2)} \quad (8)$$

[0193] また、ファクトデータの生成モデルは、以下に示す式（9）で表される。

[0194] [数14]

$$x_{d_1 d_2} \sim \text{Bern}(\theta_{k_1 k_2}^{[1]}) \quad (9)$$

[0195] また、そのパラメータの変分事後分布は、以下に示す式（10）で表される。

[0196] [数15]

$$q(\theta_{k_1 k_2}^{[1]}) = \text{Beta}(\hat{a}_{k_1 k_2}^{[1]}, \hat{b}_{k_1 k_2}^{[1]}) \quad (10)$$

[0197] また、そのパラメータの更新式は、以下に示す式（11）、式（12）で表される。

[0198]

[数16]

$$\hat{a}_{k_1 k_2}^{[1]} = a_0^{[1]} + \sum_{d_1}^{D^{(1)}} \sum_{d_2}^{D^{(2)}} \phi_{d_1, k_1}^{(1)} \phi_{d_2, k_2}^{(2)} x_{d_1 d_2} \quad (11)$$

[0199] [数17]

$$\hat{b}_{k_1 k_2}^{[1]} = b_0^{[1]} + \sum_{d_1}^{D^{(1)}} \sum_{d_2}^{D^{(2)}} \phi_{d_1, k_1}^{(1)} \phi_{d_2, k_2}^{(2)} (1 - x_{d_1 d_2}) \quad (12)$$

[0200] また、SVM (Support Vector Machine) の重みパラメータの変分事後分布は、以下に示す式 (13) で表される。

[0201] [数18]

$$q(\eta_{k_1}^{(1)}) = N(\hat{\mu}_{k_1^{(1)}}, I) \quad (13)$$

[0202] また、そのパラメータの更新式は、以下に示す式 (14) で表される。

[0203] [数19]

$$\hat{\mu}_{k_1^{(1)}} = w_0^{(1)} + \sum_{d_1}^{D^{(1)}} \phi_{d_1, k_1}^{(1)} \sum_{y_1}^{Y^{(1)}} \omega_{y_1, d_1}^{(1)} f_{d_1}^{\Delta}(y_1) \quad (14)$$

[0204] また、SVMの学習問題は、以下に示す式（１５）で表される。

[0205] [数20]

$$\begin{aligned} \max_{\omega^{(1)}} \quad & -\frac{1}{2} \sum_{k_1}^{K^{(1)}} \hat{\mu}_{k_1^{(1)}}^\top \hat{\mu}_{k_1^{(1)}} + \sum_{y_1}^{Y^{(1)}} \sum_{d_1}^{D^{(1)}} \omega_{y_1, d_1}^{(1)} l_{d_1}^\Delta(y^{(1)}) \\ \text{s.t.} \quad & 0 \leq \sum_{y_1}^{Y^{(1)}} \omega_{y_1, d_1}^{(1)} \leq C_1^{(1)} \end{aligned}$$

（１５）

[0206] なお、式（１５）において、 $\mu_{k_1}^{(1)}$ は、以下に示す式（１６）で表される。

[0207] [数21]

$$\hat{\mu}_{k_1^{(1)}} = w_0^{(1)} + \sum_{d_1}^{D^{(1)}} \phi_{d_1, k_1}^{(1)} \sum_{y_1}^{Y^{(1)}} \omega_{y_1, d_1}^{(1)} f_{d_1}^\Delta(y_1)$$

（１６）

[0208] 以下、第１の実施形態の具体例として、上記の式を用いた処理経過の例を示す。図１９、図２０は、第１の実施形態の具体例における処理経過の例を示すフローチャートである。

[0209] まず、データ入力部２がデータを取得する（ステップＳ３００）。

[0210] 次に、初期化部３１が、クラスタを初期化する（ステップＳ３０２）。

[0211] 次に、予測モデル学習部３２１は、データセット１の各クラスタで式（１５）を解き、パラメータ ω を取得する（ステップＳ３０４）。

[0212] 次に、予測モデル学習部３２１は、データセット１の各クラスタで、式（１４）によりSVMモデル $q(\eta_{k_1}^{(1)})$ を更新する（ステップＳ３０６）。

- [0213] 次に、クラスタ割り当て部322は、データセット1の各データのクラスタ割り当て $q(z_{d1}^{(1)} = k_1)$ を式(1)に従って更新する(ステップS308)。
- [0214] 次に、クラスタ割り当て部322は、データセット2の各データのクラスタ割り当て $q(z_{d2}^{(2)} = k_2)$ を式(2)に従って更新する(ステップS310)。
- [0215] 次に、クラスタ情報算出部323は、データセット1の各クラスタのモデル $q(v_{k1}^{(1)})$ を式(6)に従って更新する(ステップS316)。
- [0216] 次に、クラスタ情報算出部323は、データセット2の各クラスタのモデル $q(v_{k2}^{(2)})$ を式(8)に従って更新する(ステップS318)。
- [0217] 次に、クラスタ関係算出部324は、データセット1, 2のクラスタの組み合わせについて、クラスタの関連度 $q(\theta_{k1k2}^{[1]})$ を式(12)に従って更新する(ステップS320)。
- [0218] 次に、終了判定部325は、終了条件が満たされたか否かを判定する(ステップS322)。終了条件が満たされていないと判定された場合(ステップS322のNo)、ラスタリング部32は、ステップS304以降の処理を繰り返す。
- [0219] 終了条件が満たされたと判定した場合(ステップS322のYes)、結果出力部5は、その時点におけるクラスタリング部32による処理の結果を出力し、処理を終了する。
- [0220] 図21は、本発明の各実施形態に係るコンピュータの構成例を示す概略ブロック図である。コンピュータ1000は、CPU1001と、主記憶装置1002と、補助記憶装置1003と、インタフェース1004とを備える。
- [0221] 各実施形態のシステム(第1、第3の実施形態における共クラスタリングシステム、第2の実施形態における予測システム)は、コンピュータ1000に実装される。各実施形態のシステムの動作は、プログラムの形式で補助記憶装置1003に記憶されている。CPU1001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、そのプログラムに従って上記の処理を実行する。

[0222] 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、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 がそのプログラムを主記憶装置 1002 に展開し、上記の処理を実行してもよい。

[0223] また、プログラムは、前述の処理の一部を実現するためのものであってもよい。さらに、プログラムは、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前述の処理を実現する差分プログラムであってもよい。

[0224] また、各装置の各構成要素の一部または全部は、汎用または専用の回路（circuitry）、プロセッサ等やこれらの組み合わ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または全部は、上述した回路等とプログラムとの組み合わせによって実現されてもよい。

[0225] 各装置の各構成要素の一部また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、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。

[0226] 次に、本発明の概要について説明する。図 22 は、本発明の共クラスタリングシステムの概要を示すブロック図である。本発明の共クラスタリングシステムは、共クラスタリング手段 71 と、予測モデル生成手段 72 と、判定手段 73 とを備える。

- [0227] 共クラスタリング手段 7 1（例えば、クラスタ割り当て部 3 2 2）は、第 1 のマスタデータと、第 2 のマスタデータと、第 1 のマスタデータ内のレコードの ID である第 1 ID と第 2 のマスタデータ内のレコードの ID である第 2 ID との関係を示すファクトデータとに基づいて、第 1 ID および第 2 ID を共クラスタリングする共クラスタリング処理を実行する。
- [0228] 予測モデル生成手段 7 2（例えば、予測モデル学習部 3 2 1）は、少なくとも第 1 ID のクラスタ毎に予測モデルを生成する予測モデル生成処理を実行する。
- [0229] 判定手段 7 3（例えば、終了判定部 3 2 5）は、所定の条件が満たされたか否かを判定する。
- [0230] 共クラスタリングシステムは、所定の条件が満たされたと判定されるまで、予測モデル生成処理および共クラスタリング処理を繰り返す。
- [0231] 共クラスタリング手段 7 1 は、1 つの第 1 ID が 1 つのクラスタに属する所属確率を決定する際、第 1 ID に対応する目的変数の値を、クラスタに対応する予測モデルを用いて予測し、当該値と実際の値との差が小さいほど、所属確率を高い確率とする。
- [0232] そのような構成により、クラスタ毎の予測モデルの予測精度をより向上させることができる。
- [0233] また、目的変数が未知である新たな第 1 ID のレコードと、新たな第 1 ID と第 2 のマスタデータ内の第 2 ID との関係を示すデータとを含むテストデータが与えられたときに、その目的変数の値を予測する予測手段（例えば、図 1 8 に示す予測部 7）を備える構成であってもよい。
- [0234] 予測手段が、
新たな第 1 ID のレコードに含まれる属性の値、または、新たな第 1 ID と第 2 のマスタデータ内の第 2 ID との関係を示すデータを用いることによって、新たな第 1 ID が属するクラスタを特定し、
そのクラスタに対応する予測モデルに、新たな第 1 ID のレコードを適用することによって目的変数の値を予測する構成であってもよい。

[0235] また、予測手段が、

新たな第1 IDのレコードに含まれる属性の値、または、新たな第1 IDと第2のマスタデータ内の第2 IDとの関係を示すデータを用いることによって、新たな第1 IDが、第1 IDの各クラスタに所属する所属確率を求め、

第1 IDの各クラスタに対応する予測モデル毎に、新たな第1 IDのレコードを適用することによって目的変数の値を予測し、予測した各値に対して、新たな第1 IDが各クラスタに所属する所属確率で重み付け加算した結果を、目的変数の値として確定する構成であってもよい。

[0236] 以上、実施形態を参照して本願発明を説明したが、本願発明は上記の実施形態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。

[0237] この出願は、2016年3月16日に提出された日本特許出願2016-052737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。

産業上の利用の可能性

[0238] 本発明は、2種類の事項それぞれをクラスタリングする共クラスタリングシステムに好適に適用される。

符号の説明

- [0239]
- 1 共クラスタリングシステム
 - 2 データ入力部
 - 3 処理部
 - 4 記憶部
 - 5 結果出力部
 - 6 テストデータ入力部
 - 7 予測部
 - 8 予測結果出力部
 - 31 初期化部

3 2 クラスタリング部

3 2 1 予測モデル学習部

3 2 2 クラスタ割り当て部

3 2 3 クラスタ情報算出部

3 2 4 クラスタ関係算出部

3 2 5 終了判定部

請求の範囲

- [請求項1] 第1のマスタデータと、第2のマスタデータと、前記第1のマスタデータ内のレコードのIDである第1IDと前記第2のマスタデータ内のレコードのIDである第2IDとの関係を示すファクトデータとに基づいて、前記第1IDおよび前記第2IDを共クラスタリングする共クラスタリング処理を実行する共クラスタリング手段と、
- 少なくとも前記第1IDのクラスタ毎に予測モデルを生成する予測モデル生成処理を実行する予測モデル生成手段と、
- 所定の条件が満たされたか否かを判定する判定手段とを備え、
- 前記所定の条件が満たされたと判定されるまで、前記予測モデル生成処理および前記共クラスタリング処理を繰り返し、
- 前記共クラスタリング手段は、
- 1つの第1IDが1つのクラスタに属する所属確率を決定する際、前記第1IDに対応する目的変数の値を、前記クラスタに対応する予測モデルを用いて予測し、当該値と実際の値との差が小さいほど、前記所属確率を高い確率とする
- ことを特徴とする共クラスタリングシステム。
- [請求項2] 目的変数が未知である新たな第1IDのレコードと、前記新たな第1IDと第2のマスタデータ内の第2IDとの関係を示すデータとを含むテストデータが与えられたときに、前記目的変数の値を予測する予測手段を備える
- 請求項1に記載の共クラスタリングシステム。
- [請求項3] 予測手段は、
- 新たな第1IDのレコードに含まれる属性の値、または、前記新たな第1IDと第2のマスタデータ内の第2IDとの関係を示すデータを用いることによって、前記新たな第1IDが属するクラスタを特定し、
- 前記クラスタに対応する予測モデルに、新たな第1IDのレコード

を適用することによって目的変数の値を予測する

請求項2に記載の共クラスタリングシステム。

[請求項4]

予測手段は、

新たな第1 IDのレコードに含まれる属性の値、または、前記新たな第1 IDと第2のマスタデータ内の第2 IDとの関係を示すデータを用いることによって、前記新たな第1 IDが、第1 IDの各クラスタに所属する所属確率を求め、

第1 IDの各クラスタに対応する予測モデル毎に、新たな第1 IDのレコードを適用することによって目的変数の値を予測し、予測した各値に対して、前記新たな第1 IDが各クラスタに所属する所属確率で重み付け加算した結果を、前記目的変数の値として確定する

請求項2に記載の共クラスタリングシステム。

[請求項5]

第1のマスタデータと、第2のマスタデータと、前記第1のマスタデータ内のレコードのIDである第1 IDと前記第2のマスタデータ内のレコードのIDである第2 IDとの関係を示すファクトデータとに基づいて、前記第1 IDおよび前記第2 IDを共クラスタリングする共クラスタリング処理を実行し、

少なくとも前記第1 IDのクラスタ毎に予測モデルを生成する予測モデル生成処理を実行し、

所定の条件が満たされたか否かを判定し、

前記所定の条件が満たされたと判定されるまで、前記予測モデル生成処理および前記共クラスタリング処理を繰り返し、

前記共クラスタリング処理で、

1つの第1 IDが1つのクラスタに属する所属確率を決定する際、前記第1 IDに対応する目的変数の値を、前記クラスタに対応する予測モデルを用いて予測し、当該値と実際の値との差が小さいほど、前記所属確率を高い確率とする

ことを特徴とする共クラスタリング方法。

[請求項6]

コンピュータに、

第1のマスタデータと、第2のマスタデータと、前記第1のマスタデータ内のレコードのIDである第1IDと前記第2のマスタデータ内のレコードのIDである第2IDとの関係を示すファクトデータとに基づいて、前記第1IDおよび前記第2IDを共クラスタリングする共クラスタリング処理、

少なくとも前記第1IDのクラス毎に予測モデルを生成する予測モデル生成処理、および、

所定の条件が満たされたか否かを判定する判定処理を実行させ、

前記所定の条件が満たされたと判定されるまで、前記予測モデル生成処理および前記共クラスタリング処理を繰り返させ、

前記共クラスタリング処理で、

1つの第1IDが1つのクラスに属する所属確率を決定する際、前記第1IDに対応する目的変数の値を、前記クラスに対応する予測モデルを用いて予測させ、当該値と実際の値との差が小さいほど、前記所属確率を高い確率とさせる

ための共クラスタリングプログラム。

[図1]

顧客ID	年齢	年収	年間のエステティック サロンの利用回数
顧客1	45	700	15
顧客2	30	500	1
顧客3	27	400	?
顧客4	50	900	?

[図2]

商品ID	商品名	価格
商品1	炭酸水	150
商品2	焼酎	900
商品3	アサイー	200
商品4	おつまみ	200

[図3]

	商品1	商品2	商品3	商品4
顧客1	1	0	1	0
顧客2	1	1	0	1
顧客3	1	1	1	1
顧客4	1	0	1	0

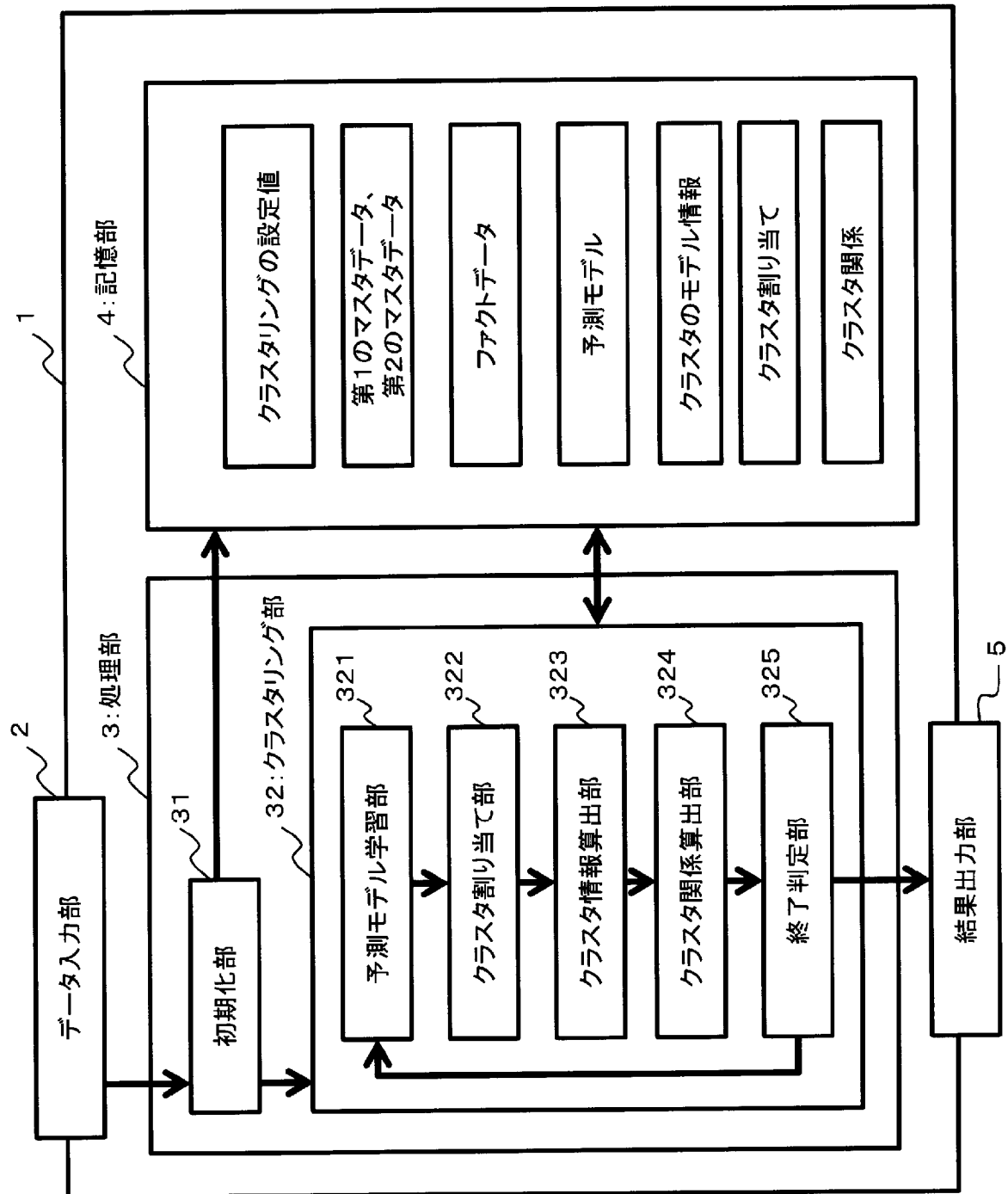
[図4]

	クラスタ番号
顧客1	3
顧客2	1
顧客3	2
顧客4	2

[図5]

	顧客クラスタ1	顧客クラスタ2	顧客クラスタ3
顧客1	0.3	0.1	0.6
顧客2	0.7	0.2	0.1
顧客3	0.1	0.9	0.1
顧客4	0.2	0.7	0.1

[図6]



[図7]

	年 齢	年収	年間のエステティック サロンの利用回数
顧客1	45	700	15
顧客2	30	500	1
顧客3	27	400	20
予測対象 顧客	50	900	?

[図8]

顧客IDクラスタ1 (美容好き)	商品IDクラスタ1 (美容商品)	商品IDクラスタ2 (酒)
顧客IDクラスタ2 (酒好き)	0.8	0.35
	0.1	0.9

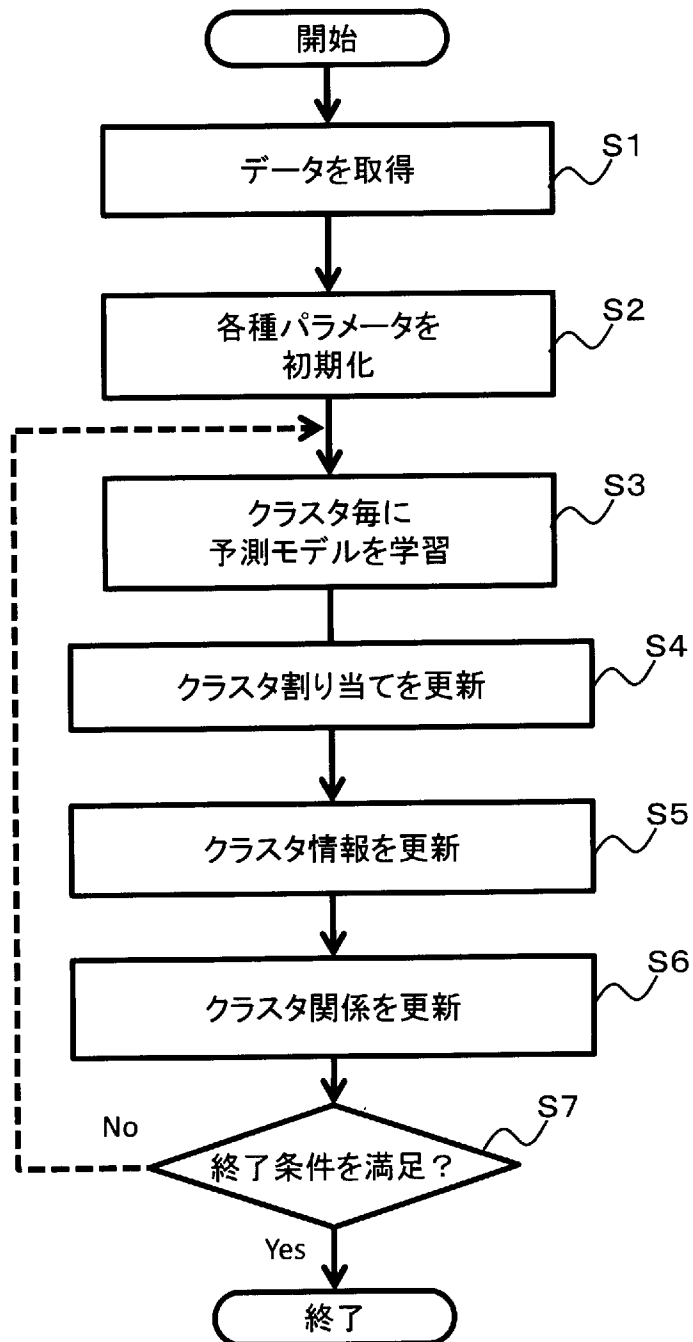
[図9]

	$k_2=1$ (甘い)	$k_2=2$ (辛い)	$k_2=3$ (苦い)
$k_1=1$ (甘党)	0.9	0.3	0.1
$k_1=2$ (辛党)	0.1	0.8	0.3

[図10]

	$j=1$	$j=2$	$j=3$
$i=1$	1	0	1
$i=2$	0	1	0
$i=3$	1	0	0
	$\vdots$	$\vdots$	$\vdots$

[図11]



[図12]

顧客ID	年齢	年収	年間のエステティック サロンの利用回数	炭酸水	価格	焼酎	価格	...
顧客1	45	700	15	1	150	0	900	...
顧客2	30	500	1	1	150	1	900	...
顧客3	27	400	?	1	150	1	900	...
顧客4	50	900	?	1	150	0	900	...

[図13]

顧客ID	年齢	年収	...	年間のエステティック サロンの利用回数
顧客1	45	700	...	15
顧客2	30	500	...	1
顧客3	27	400	...	?
顧客4	50	900	...	?

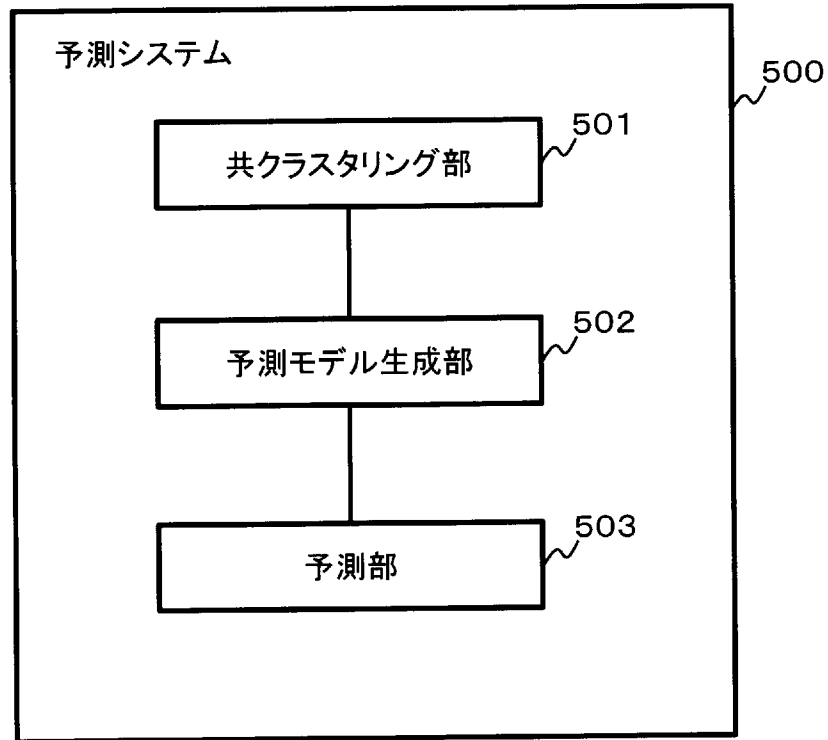
[図14]

商品ID	価格	...
菓子1	10	...
炭酸飲料P	130	...
炭酸飲料Q	130	...

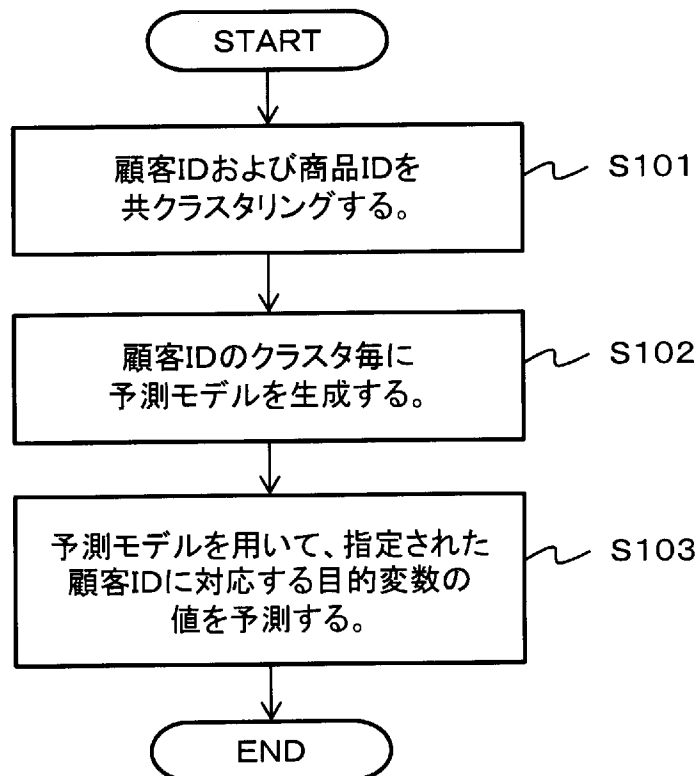
[図15]

	菓子1	炭酸飲料P	炭酸飲料Q
顧客1	0	1	0
顧客2	1	1	0
顧客3	0	1	0
顧客4	1	1	4
:	:	:	:

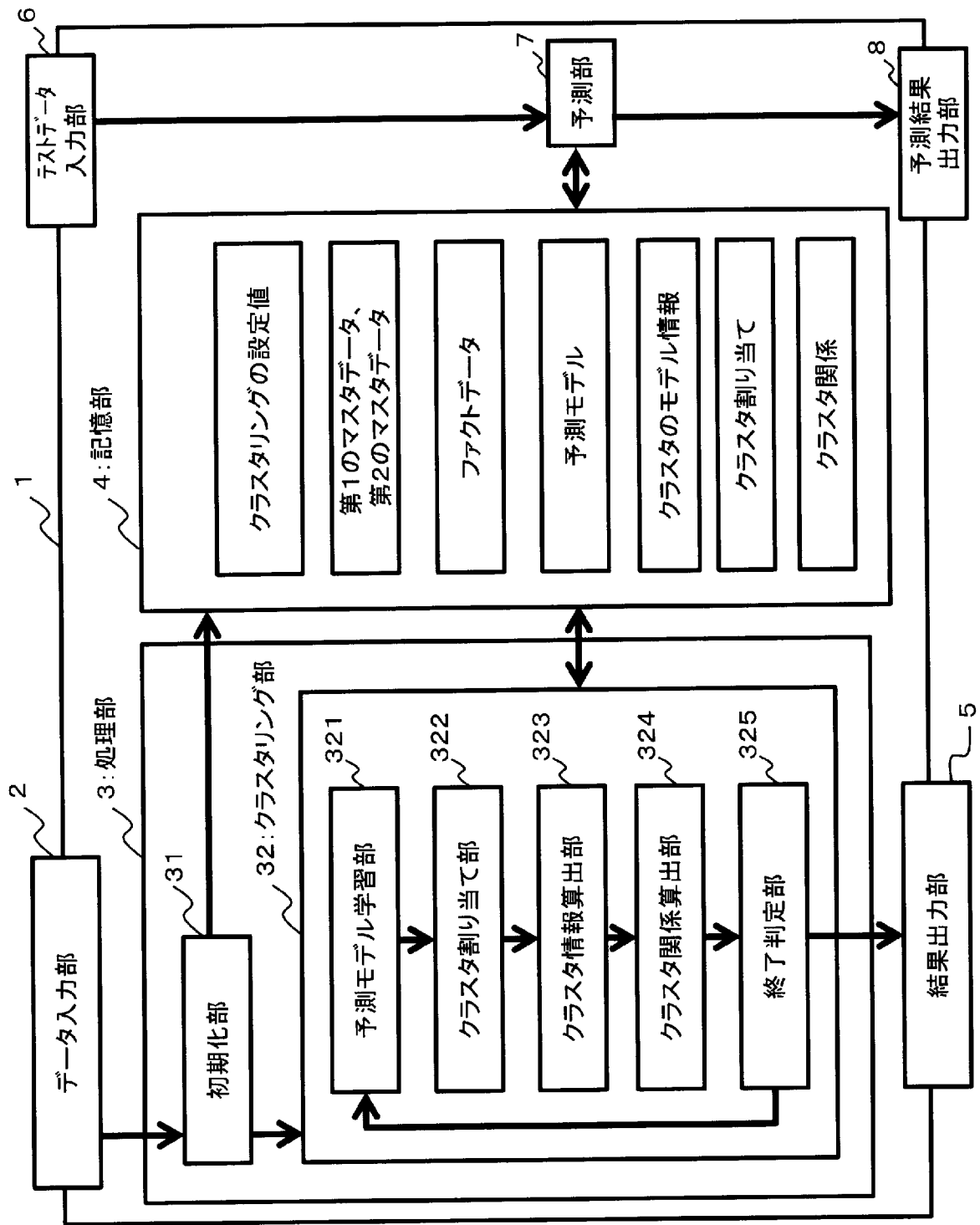
[図16]



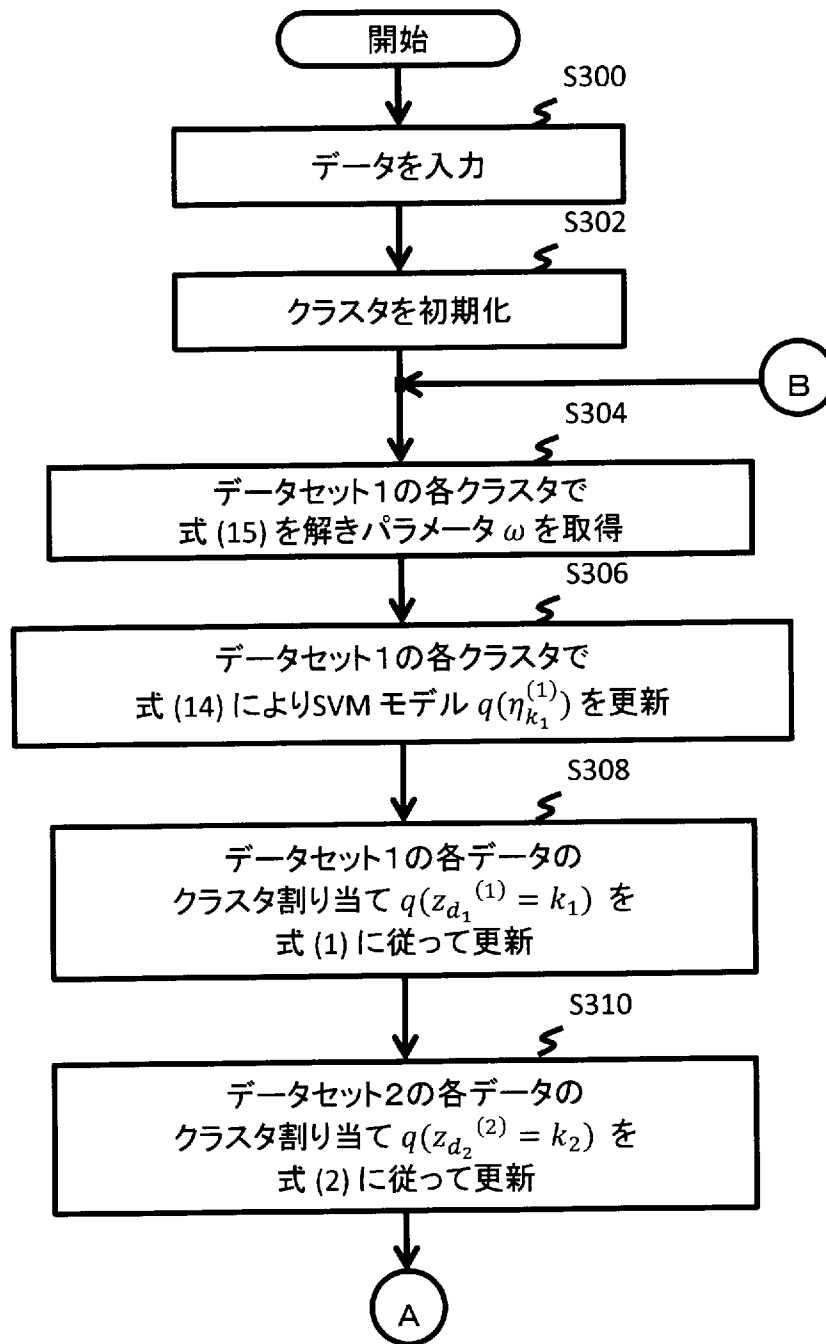
[図17]



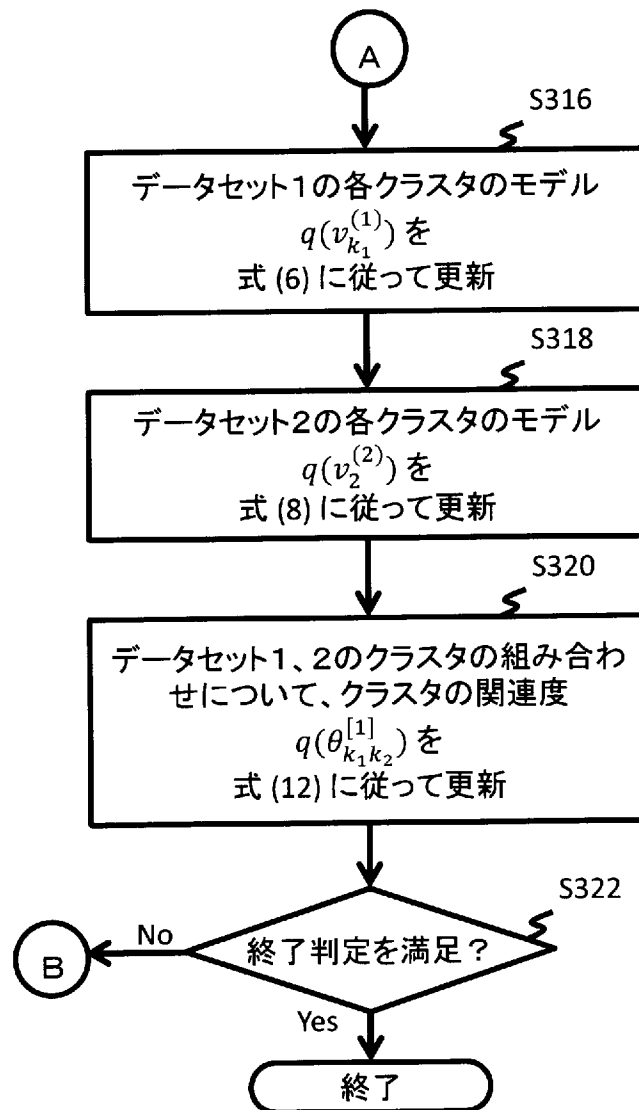
[図18]



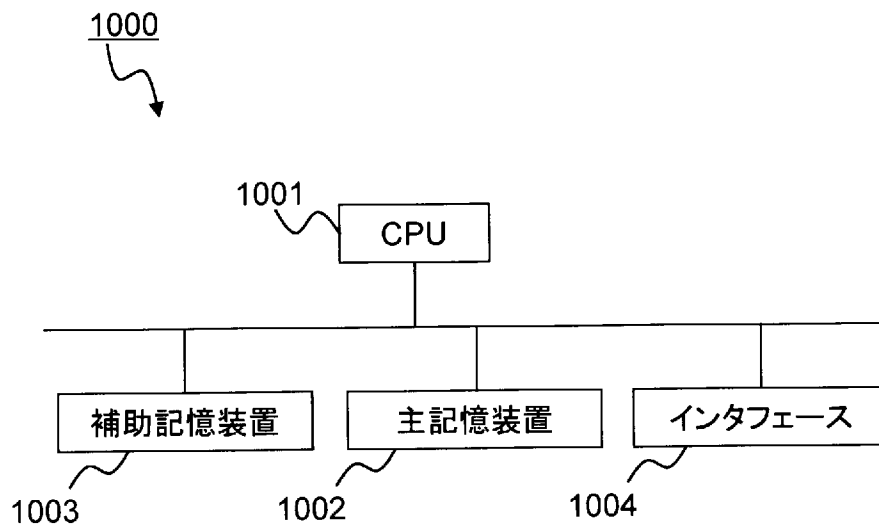
[図19]



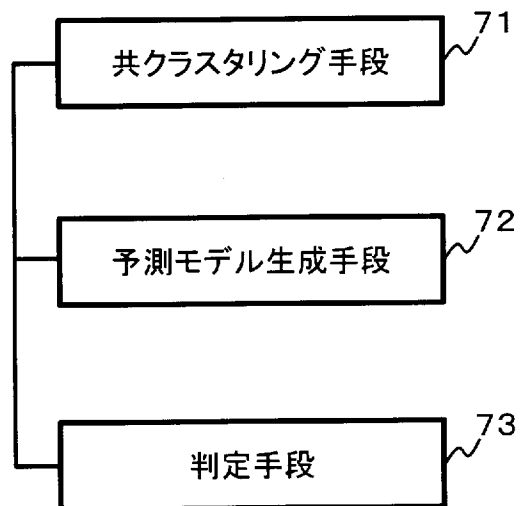
[図20]



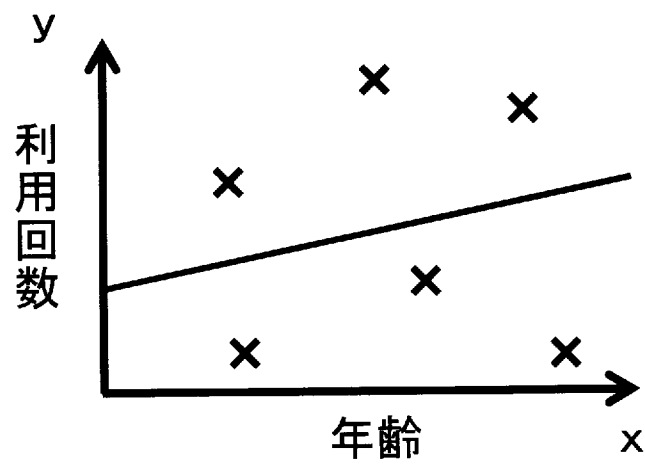
[図21]



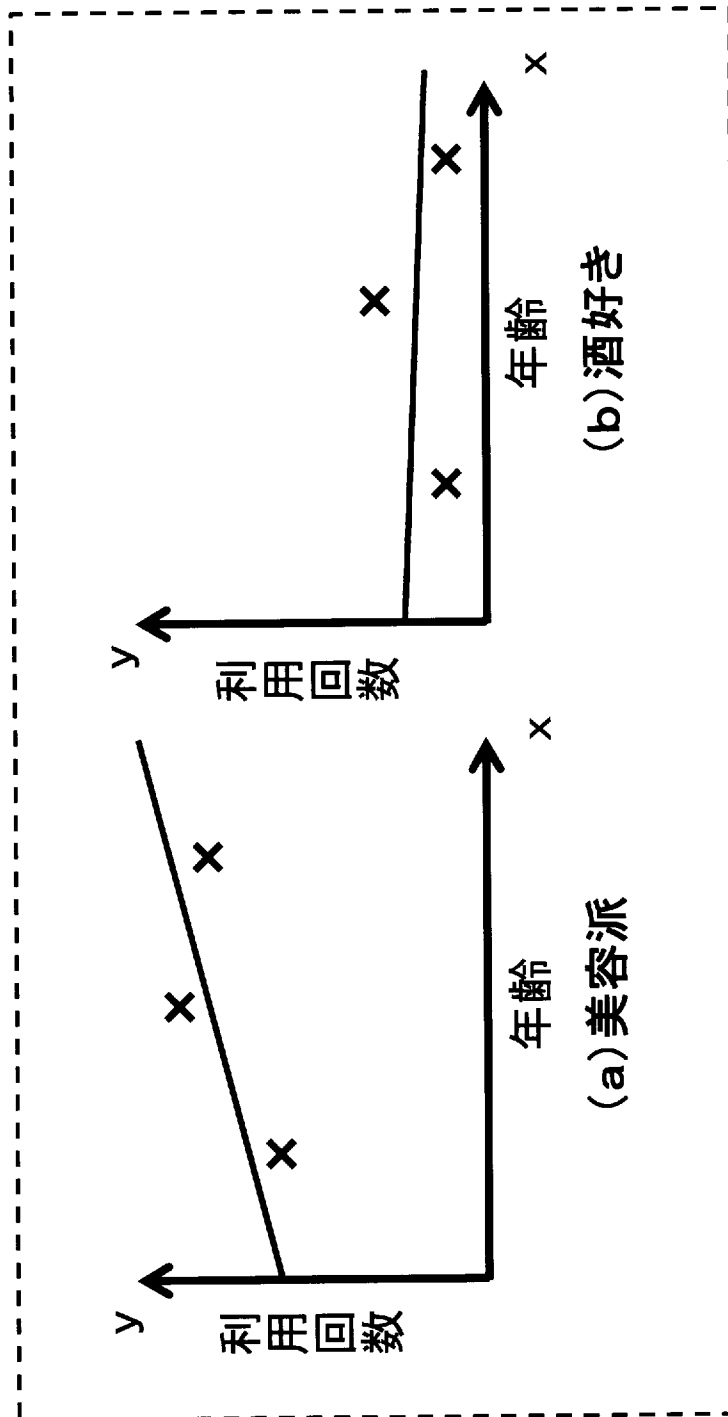
[図22]



[図23]



[図24]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2017/008488

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/30(2006.01)i, G06N99/00(2010.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/30, G06N99/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2017
Kokai Jitsuyo Shinan Koho	1971-2017	Toroku Jitsuyo Shinan Koho	1994-2017

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2014/179724 A1 (NEW YORK UNIVERSITY), 06 November 2014 (06.11.2014), paragraphs [0025] to [0038]; fig. 2 & US 2016/0110730 A1	1-6
A	US 2009/0055139 A1 (AGARWAL, Deepak), 26 February 2009 (26.02.2009), entire text; all drawings (Family: none)	1-6
A	JP 2007-164346 A (Toshiba Corp.), 28 June 2007 (28.06.2007), entire text; all drawings (Family: none)	1-6

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
07 April 2017 (07.04.17)

Date of mailing of the international search report
25 April 2017 (25.04.17)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2017/008488

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
P, X	Masafumi OYAMADA, Shinji NAKADAI, "On Modeling Relational Infinite SVM", 2016 Nendo Annual Conference of JSAI (JSAI2016) Ronbunshu, 06 June 2016 (06.06.2016), 1K2-1in1, pages 1 to 4, Internet:<URL: http://kaigi.org/jsai/webprogram/2016/paper-310.html >	1-6

A. 発明の属する分野の分類 (国際特許分類 (IPC))

Int.Cl. G06F17/30(2006.01)i, G06N99/00(2010.01)i

B. 調査を行った分野

調査を行った最小限資料 (国際特許分類 (IPC))

Int.Cl. G06F17/30, G06N99/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2017年
日本国実用新案登録公報	1996-2017年
日本国登録実用新案公報	1994-2017年

国際調査で使用した電子データベース (データベースの名称、調査に使用した用語)

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	WO 2014/179724 A1 (NEW YORK UNIVERSITY) 2014.11.06, [0025]-[0038], Figure 2 & US 2016/0110730 A1	1-6
A	US 2009/0055139 A1 (AGARWAL, Deepak) 2009.02.26, 全文, 全図 (ファミリーなし)	1-6

☒ C欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」 特に関連のある文献ではなく、一般的技術水準を示すもの

「E」 国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」 優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献 (理由を付す)

「O」 口頭による開示、使用、展示等に言及する文献

「P」 国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」 国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」 特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」 特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」 同一パテントファミリー文献

国際調査を完了した日

07.04.2017

国際調査報告の発送日

25.04.2017

国際調査機関の名称及びあて先

日本国特許庁 (ISA/J P)

郵便番号 100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官 (権限のある職員)

小太刀 慶明

5M

6291

電話番号 03-3581-1101 内線 3599

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 2007-164346 A (株式会社東芝) 2007.06.28, 全文, 全図 (ファミリーなし)	1-6
P, X	小山田 昌史, 中台 慎二, “無限混合SVMの関係モデル化”, 2016年度人工知能学会全国大会 (JSAI 2016) 論文集, 2016.06.06, 1 K 2 - 1 i n 1, p. 1 - 4, インターネット: <URL: http://kaigi.org/jsai/webprogram/2016/paper-310.html >	1-6