(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5391637号
(P5391637)
(45)発行日　平成26年1月15日(2014.1.15)　(24)登録日　平成25年10月25日(2013.10.25)

(51)Int.Cl.　F I
G06F 17/30 (2006.01)　G06F 17/30 210D
G06F 17/30 350C

請求項の数 24　(全 27 頁)

(21)出願番号　特願2008-264442(P2008-264442)
(22)出願日　平成20年10月10日(2008.10.10)
(65)公開番号　特開2010-92432(P2010-92432A)
(43)公開日　平成22年4月22日(2010.4.22)
審査請求日　平成23年9月8日(2011.9.8)

(73)特許権者　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(74)代理人　100103090
弁理士　岩壁　冬樹
(74)代理人　100124501
弁理士　塩川　誠人
(72)発明者　黒岩　由希子
東京都港区芝五丁目７番１号　日本電気株式会社内

審査官　野崎　大進

最終頁に続く

(54)【発明の名称】データ類似度計算システム、データ類似度計算方法およびデータ類似度計算プログラム

(57)【特許請求の範囲】
【請求項１】
データの特徴を示す属性とデータの類別を示すクラスとを含むデータの集合から、データの重複を許して前記集合の部分集合を複数回生成する部分集合生成手段と、
部分集合が生成される毎に、属性からクラスを判定するルールである分類器を、前記部分集合に基づいて生成する分類器生成手段と、
分類器が生成される毎に、分類器を用いて、前記部分集合に属する個々のデータのクラスを判定するクラス判定手段と、
データの集合の部分集合が生成され、クラス判定手段が当該部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ同士の類似度に値を加算する類似度算出手段とを備える
ことを特徴とするデータ類似度計算システム。
【請求項２】
データ同士の類似度に基づいて、データの集合に属するデータを複数のグループに分類するデータグループ化手段を備える
請求項１に記載のデータ類似度計算システム。
【請求項３】
データグループ化手段は、データの集合に属する個々のデータをそれぞれ別々のグループに分類し、互いに異なる二つのグループに属するデータ同士の類似度を求め、前記類似度が最大となる二つのグループを併合することを繰り返し、グループの総数を目標数まで

減少させる
請求項２に記載のデータ類似度計算システム。
【請求項４】
データ集合に属するデータを、特定の属性の属性値に応じて、グループに分類する属性データ分類手段と、
データグループ化手段によって分類されたデータのグループと、属性データ分類手段によって分類されたデータのグループとの関係に基づいて、類似度算出に対する前記特定の属性の関連度を計算する関連度計算手段とを備える
請求項２または請求項３に記載のデータ類似度計算システム。
【請求項５】
部分集合生成手段は、データの集合からデータをランダムサンプリングすることによって、前記集合の部分集合を生成する
請求項１から請求項４のうちのいずれか１項に記載のデータ類似度計算システム。
【請求項６】
類似度算出手段は、クラス判定手段によって特定のクラスと判定されたデータ同士の類似度に対してのみ値を加算する
請求項１から請求項５のうちのいずれか１項に記載のデータ類似度計算システム。
【請求項７】
分類器生成手段は、部分集合に属するデータのうち所与のクラスが特定のクラスであるデータを加重して分類器を生成する
請求項１から請求項６のうちのいずれか１項に記載のデータ類似度計算システム。
【請求項８】
部分集合生成手段は、少なくとも顧客の特徴または販売条件を属性とし顧客の行動をクラスとするデータの集合から、当該集合の部分集合を生成する
請求項１から請求項７のうちのいずれか１項に記載のデータ類似度計算システム。
【請求項９】
部分集合生成手段が、データの特徴を示す属性とデータの類別を示すクラスとを含むデータの集合から、データの重複を許して前記集合の部分集合を複数回生成する部分集合生成ステップと、
分類器生成手段が、部分集合が生成される毎に、属性からクラスを判定するルールである分類器を、前記部分集合に基づいて生成する分類器生成ステップと、
クラス判定手段が、分類器が生成される毎に、分類器を用いて、前記部分集合に属する個々のデータのクラスを判定するクラス判定ステップと、
類似度算出手段が、データの集合の部分集合が生成され、クラス判定ステップで当該部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ同士の類似度に値を加算する類似度算出ステップとを含む
ことを特徴とするデータ類似度計算方法。
【請求項１０】
データグループ化手段が、データ同士の類似度に基づいて、データの集合に属するデータを複数のグループに分類するデータグループ化ステップを含む
請求項９に記載のデータ類似度計算方法。
【請求項１１】
データグループ化手段が、データグループ化ステップで、データの集合に属する個々のデータをそれぞれ別々のグループに分類し、互いに異なる二つのグループに属するデータ同士の類似度を求め、前記類似度が最大となる二つのグループを併合することを繰り返し、グループの総数を目標数まで減少させる
請求項１０に記載のデータ類似度計算方法。
【請求項１２】
属性データ分類手段が、データ集合に属するデータを、特定の属性の属性値に応じて、グループに分類する属性データ分類ステップと、

関連度計算手段が、データグループ化ステップで分類されたデータのグループと、属性データ分類ステップで分類されたデータのグループとの関係に基づいて、類似度算出に対する前記特定の属性の関連度を計算する関連度計算ステップとを備える
請求項１０または請求項１１に記載のデータ類似度計算方法。
【請求項１３】
部分集合生成手段が、部分集合生成ステップで、データの集合からデータをランダムサンプリングすることによって、前記集合の部分集合を生成する
請求項９から請求項１２のうちのいずれか１項に記載のデータ類似度計算方法。
【請求項１４】
類似度算出手段が、類似度算出ステップで、クラス判定ステップで特定のクラスと判定されたデータ同士の類似度に対してのみ値を加算する
請求項９から請求項１３のうちのいずれか１項に記載のデータ類似度計算方法。
【請求項１５】
分類器生成手段が、分類器生成ステップで、部分集合に属するデータのうち所与のクラスが特定のクラスであるデータを加重して分類器を生成する
請求項９から請求項１４のうちのいずれか１項に記載のデータ類似度計算方法。
【請求項１６】
部分集合生成手段が、部分集合生成ステップで、少なくとも顧客の特徴または販売条件を属性とし顧客の行動をクラスとするデータの集合から、当該集合の部分集合を生成する
請求項９から請求項１５のうちのいずれか１項に記載のデータ類似度計算方法。
【請求項１７】
コンピュータに、
データの特徴を示す属性とデータの類別を示すクラスとを含むデータの集合から、データの重複を許して前記集合の部分集合を複数回生成する部分集合生成処理、
部分集合が生成される毎に、属性からクラスを判定するルールである分類器を、前記部分集合に基づいて生成する分類器生成処理、
分類器が生成される毎に、分類器を用いて、前記部分集合に属する個々のデータのクラスを判定するクラス判定処理、および、
データの集合の部分集合が生成され、クラス判定処理で当該部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ同士の類似度に値を加算する類似度算出処理
を実行させるためのデータ類似度計算プログラム。
【請求項１８】
コンピュータに、
データ同士の類似度に基づいて、データの集合に属するデータを複数のグループに分類するデータグループ化処理
を実行させる請求項１７に記載のデータ類似度計算プログラム。
【請求項１９】
コンピュータに、
データグループ化処理で、データの集合に属する個々のデータをそれぞれ別々のグループに分類させ、互いに異なる二つのグループに属するデータ同士の類似度を求めさせ、前記類似度が最大となる二つのグループを併合することを繰り返させ、グループの総数を目標数まで減少させる
請求項１８に記載のデータ類似度計算プログラム。
【請求項２０】
コンピュータに、
データ集合に属するデータを、特定の属性の属性値に応じて、グループに分類する属性データ分類処理、および、
データグループ化処理で分類されたデータのグループと、属性データ分類処理で分類されたデータのグループとの関係に基づいて、類似度算出に対する前記特定の属性の関連度

を計算する関連度計算処理
を実行させる請求項１８または請求項１９に記載のデータ類似度計算プログラム。
【請求項２１】
コンピュータに、
部分集合生成処理で、データの集合からデータをランダムサンプリングすることによって、前記集合の部分集合を生成させる
請求項１７から請求項２０のうちのいずれか１項に記載のデータ類似度計算プログラム。
【請求項２２】
コンピュータに、
類似度算出処理で、クラス判定処理で特定のクラスと判定されたデータ同士の類似度に対してのみ値を加算させる
請求項１７から請求項２１のうちのいずれか１項に記載のデータ類似度計算プログラム。
【請求項２３】
コンピュータに、
分類器生成処理で、部分集合に属するデータのうち所与のクラスが特定のクラスであるデータを加重して分類器を生成させる
請求項１７から請求項２２のうちのいずれか１項に記載のデータ類似度計算プログラム。
【請求項２４】
コンピュータに、
部分集合生成処理で、少なくとも顧客の特徴または販売条件を属性とし顧客の行動をクラスとするデータの集合から、当該集合の部分集合を生成させる
請求項１７から請求項２３のうちのいずれか１項に記載のデータ類似度計算プログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、データ間の類似度を計算するデータ類似度計算システム、データ類似度計算方法およびデータ類似度計算プログラムに関する。
【背景技術】
【０００２】
小売店では、ＰＯＳ（Ｐｏｉｎｔ　ｏｆ　Ｓａｌｅｓ）システムが用いられており、収集された商品の売上データは様々に活用されている。そして、顧客の特徴に基づいて顧客を分類して、売上データを用いて種々の分析を行うシステムが提案されている。例えば、特許文献１には、分析者の意志に従って顧客を分類し、顧客の購買履歴状況を確認可能とするシステムが記載されている。
【０００３】
また、特許文献２には、ＰＯＳデータを用いて消費者の併売傾向を把握するシステムが記載されている。
【０００４】
また、特許文献３には、各種カテゴリにおける複数の商品を関連づけて登録するための複数の仮想ＢＯＸを有するデータベースを備え、顧客に対する推奨商品グループを提示できるシステムが記載されている。
【０００５】
また、データの類似度を求める装置が特許文献４に記載されている。特許文献４に記載された装置は、予め保持されている事例と、与えられた事例（クエリ）との類似度である事例間類似度を計算する。また、保持されている事例（例えば、ある地点の気象データ）にはクラス（例えば、気象データ計測時点から３時間後の天気）が定められている。特許

文献４に記載された装置は、保持されている事例から相関ルールを導出する。相関ルールは、条件（事例）に応じた結論（クラス）を導くルールである。そして、相関ルールを用いて、事例間類似度から総合類似度を計算する。

【０００６】

特許文献４には、「A_1，A_2，・・・，A_k→Ｂ」という形式で表される相関ルールが記載されている。そして、「A_1，A_2，・・・，A_k，Ｂ」はアイテム集合と称されている。また、アイテム集合A_1，A_2，・・・，A_k，Ｂを含むレコード数の全レコードに対する比率を、相関ルールの支持度と呼び、アイテム集合A_1，A_2，・・・，A_kを含むレコードの中で、アイテムＢを含むレコードの割合を相関ルールの確信度と呼ぶ。特許文献４に記載された装置は、支持度および相関度が予め定められた各々の下限値以上となるような相関ルールを抽出する。

【０００７】

また、自動的にデータを分類する装置が特許文献５に記載されている。特許文献５に記載されたデータ分類装置は、注目パターンデータとの類似度の高い順に複数のパターンデータを取りだし、近傍クラスタを決定する。

【０００８】

また、ルールを用いて予測を行う装置として、特許文献６に記載された装置がある。特許文献６に記載された装置は、ある地域（第１の地域）に生じた事象から論理規則を発生し、その規則を他の地域（第２の地域）に適用して、第２の地域でその事象が生じる傾向を予測する。

【０００９】

【特許文献１】特開平９－１０１９８４号公報（段落００１４－００５０）
【特許文献２】特開２００７－９４５９２号公報（段落００２４）
【特許文献３】特開２００３－３３７８８６号公報（段落００１０，００１５，００１６）
【特許文献４】特開２００２－１４９６９７号公報（段落００１８－００６５）
【特許文献５】特開２００３－２５６４４３号公報（段落００９０－００９３）
【特許文献６】特開２００４－１２６７５７号公報（段落００３３，００５１）

【発明の開示】

【発明が解決しようとする課題】

【００１０】

特許文献１に記載されたシステムでは、システムのユーザである分析者が、手動で顧客を分類する必要があった。そのため、顧客の購買行動から顧客の分類を自動的に行うことができなかった。また、特許文献３に記載された装置においても、各種カテゴリにおける複数の商品の関連付けを、顧客自身や店員等の商品観察者が行う必要があった。

【００１１】

データの分類を行うために、データ間の類似度を計算し、その類似度を用いてデータを分類することが考えられる。特許文献４に記載された装置は、相関ルールを導出し、総合類似度を計算する。しかし、特許文献４に記載された装置にように相関ルールを導出すると、関連のある属性（特許文献４におけるアイテム）がデータ中に多く存在する場合に、それらの関連のある属性の影響を強く受けてしまい、適切な類似度計算を行えない。

【００１２】

例えば、図１２に示すデータ番号１～１０の１０個のデータが与えられているとする。図１２に示すデータにおいて、データ番号１～４およびデータ番号６は、属性ｂ，ｃ，ｄ，ｅの値が共通であり、データの類別を表すクラスも共通である。従って、データの属性値を条件部としてクラスを結論部とする相関ルールを抽出する場合、多くのデータ間で属性値の組み合わせが共通となっている属性ｂ，ｃ，ｄ，ｅの影響を強く受けてしまう。この結果、類似度も、関連のある特定の属性に影響されてしまう。すなわち、ある属性の属性値がある値に合致しているデータ間の類似度が比較的高くなり、合致していないデータ間の類似度は比較的低くなる傾向が出てしまう。

【００１３】
そこで、本発明は、データを分類するためにデータ間の類似度を求める際に、データ中に互いに関連のある属性が多く存在してもそれらの関連のある属性の影響を強く受けることなく類似度を求めることができるデータ類似度計算システム、データ類似度計算方法およびデータ類似度計算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】
【００１４】
本発明によるデータ類似度計算システムは、データの特徴を示す属性とデータの類別を示すクラスとを含むデータの集合から、データの重複を許してその集合の部分集合を複数回生成する部分集合生成手段と、部分集合が生成される毎に、属性からクラスを判定するルールである分類器を、部分集合に基づいて生成する分類器生成手段と、分類器が生成される毎に、分類器を用いて、部分集合に属する個々のデータのクラスを判定するクラス判定手段と、データの集合の部分集合が生成され、クラス判定手段がその部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ同士の類似度に値を加算する類似度算出手段とを備えることを特徴とする。
【００１５】
本発明によるデータ類似度計算方法は、部分集合生成手段が、データの特徴を示す属性とデータの類別を示すクラスとを含むデータの集合から、データの重複を許してその集合の部分集合を複数回生成する部分集合生成ステップと、分類器生成手段が、部分集合が生成される毎に、属性からクラスを判定するルールである分類器を、部分集合に基づいて生成する分類器生成ステップと、クラス判定手段が、分類器が生成される毎に、分類器を用いて、部分集合に属する個々のデータのクラスを判定するクラス判定ステップと、類似度算出手段が、データの集合の部分集合が生成され、クラス判定ステップでその部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ同士の類似度に値を加算する類似度算出ステップとを含むことを特徴とする。
【００１６】
本発明によるデータ類似度計算プログラムは、コンピュータに、データの特徴を示す属性とデータの類別を示すクラスとを含むデータの集合から、データの重複を許してその集合の部分集合を複数回生成する部分集合生成処理、部分集合が生成される毎に、属性からクラスを判定するルールである分類器を、部分集合に基づいて生成する分類器生成処理、分類器が生成される毎に、分類器を用いて、部分集合に属する個々のデータのクラスを判定するクラス判定処理、および、データの集合の部分集合が生成され、クラス判定処理でその部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ同士の類似度に値を加算する類似度算出処理を実行させることを特徴とする。
【発明の効果】
【００１７】
本発明によれば、データを分類するためにデータ間の類似度を求める際に、データ中に互いに関連のある属性が多く存在してもそれらの関連のある属性の影響を強く受けることなく類似度を求めることができる。
【発明を実施するための最良の形態】
【００１８】
以下、本発明の実施形態を図面を参照して説明する。
ここでは、商品やサービスの顧客に関するデータ間の類似度を求める場合を例にして本発明の実施形態を説明する。なお、有料で商品やサービスを利用する者だけでなく、無料で商品やサービスを利用する者や、今後商品やサービスを利用すると思われる人も顧客と呼ぶ。無料の場合は、売価は０とする。また、商品またはサービスの概念としては、製品等の完成品だけでなく、製品の機能を維持するための最小単位も含まれる。さらに、ある商品またはサービスのカテゴリを１つの商品と考えて適用することもできる。
【００１９】
実施形態１．

図１は、本発明の第１の実施形態の例を示すブロック図である。本発明によるデータ類似度計算システム１は、部分集合生成部１１と、分類器生成部１２と、自己評価部１３と、類似度算出部１４とを備える。

【００２０】

部分集合生成部１１は、データの集合から、その集合の部分集合を生成する。データの集合は、例えば、データ類似度計算システム１に設けられたキーボード等の入力装置（図示せず）を介して入力されればよいが、データの集合の入力態様は特に限定されない。以下、データの集合とは、データ全体の集合を意味する。

【００２１】

データの集合に含まれる個々のデータは、データの特徴を示す属性と、データの類別を示すクラスとを含む。ここでは、少なくとも顧客の特徴または販売条件をデータの特徴とし、顧客の行動をクラスとする場合を例にして説明する。図２は、データの例を示す説明図である。図２に示す例では、「天気」、「気温」、「湿度」、「風が強いか否か」を属性とする場合を示している。これらの属性は、販売条件であるが、「年齢」、「性別」等の顧客の特徴を属性としてもよく、顧客の特徴および販売条件がいずれも属性となっていてもよい。また、図２に示す例では、顧客の行動は、「購入する」と「購入しない」の２種類であるものとし、データも「購入する」および「購入しない」という２種類のクラスに類別される場合を例にする。図２に例示するデータは、属性が示す条件の下で、ある商品またはサービスを顧客が購入したか否かを表している。例えば、図１に示す１番目のデータは、「晴れ」、気温「２９度」、湿度「８５％」、「風が強くない」という条件の下で、顧客が購入しなかったということを表している。

【００２２】

部分集合生成部１１は、このようなデータの集合から一部のデータを選択することで、データの集合の部分集合を生成する。また、部分集合生成部１１は、部分集合を複数回生成する。このとき、データの重複を許して部分集合を生成する。すなわち、部分集合生成部１１によって生成されたある部分集合と別の部分集合に、同じデータが属していてもよい。

【００２３】

分類器生成部１２は、データの集合の部分集合が生成される毎に、その部分集合に基づいて分類器を生成する。分類器は、属性からクラスを判定するルールである。部分集合と、その部分集合に基づいて生成された分類器とが対応する。

【００２４】

自己評価部１３は、分類器が生成される毎に、生成された分類器を用いて、その分類器に対応する部分集合に属する個々のデータのクラスを判定する。この判定は、分類器を用いて、個々のデータに含まれる属性からクラスを予測する処理であるということもできる。なお、自己評価部１３が判定したクラス（換言すれば、予測したクラス）と、データにおいて予め定められたクラスとが合致するとは限らない。

【００２５】

類似度算出部１４は、自己評価部１３で判定されたクラスに基づいて、部分集合に属する個々のデータ間の類似度を算出する。データの集合に属する任意のデータ同士の組み合わせに対して、予め類似度の初期値０が設定される。類似度算出部１４は、部分集合生成部１１によって部分集合が生成され、自己評価部１３がその部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ間の類似度に値を加算していくことによって、各データ間の類似度を算出する。

【００２６】

部分集合生成部１１、分類器生成部１２、自己評価部１３および類似度算出部１４は、例えば、プログラム（データ類似度計算プログラム）に従って動作するＣＰＵによって実現される。その場合、プログラムは、例えばデータ類似度計算システム１が備えるプログラム記憶装置（図示せず）に記憶され、ＣＰＵがプログラムを読み込み、そのプログラムに従って部分集合生成部１１、分類器生成部１２、自己評価部１３および類似度算出部１

４として動作すればよい。

【００２７】

次に、動作について説明する。

図３は、第１の実施形態のデータ類似度計算システムの処理経過の例を示すフローチャートである。例えば、データ類似度計算システム１に設けられたキーボード等の入力装置（図示せず）を介して、データの集合が入力されると、データ類似度計算システムは以下のように動作する。ただし、データの集合に属する個々のデータ（個別データと称する）の数をＮとする。また、データの集合の部分集合を生成する回数をＴとする。Ｔは、例えば、１００、５００、１０００等の数であるが、Ｔはこれらの値に限定されない。また、部分集合生成部１１が１つの部分集合に属する個別データとしてデータの集合から選択する個別データの数をＭとする。Ｍは、例えば、Ｎの１％、５％、または１０％等の値とすればよいが、Ｍはこれらの値に限定されない。部分集合の生成を繰り返す繰り返し回数であるＴや、部分集合に含める個別データ数であるＭは、それぞれ、キーボード等の入力装置（図示せず）を介して、データ類似度計算システムのユーザによって入力されてもよい。あるいは、他の態様でＴ，Ｍが指定されてもよい。

【００２８】

また、データの集合に属する個別データを順番に指定するための第１の変数をｉとし、第２の変数をｊとする。ｉは、１≦ｉ≦Ｎを満たす整数であり、同様に、ｊは、１≦ｊ≦Ｎを満たす整数である。ｉ，ｊを指定することでそれぞれ個別データを指定することができ、ｉ，ｊの組み合わせによって一対の個別データの組を指定することができる。ｉ，ｊによって指定される１対の個別データの類似度をＳｉｍ（ｉ，ｊ）と記す。

【００２９】

最初に、部分集合生成部１１は、１≦ｉ≦Ｎの範囲のｉと、１≦ｊ≦Ｎの範囲のｊによって定められるｉとｊの組み合わせを順次定め、その各組み合わせに応じた各類似度Ｓｉｍ（ｉ，ｊ）を０に初期化する（ステップＡ１）。すなわち、ｉ番目の個別データとｊ番目の個別データとの類似度となるＳｉｍ（ｉ，ｊ）を、ｉ，ｊの組み合わせ毎に０に初期化する。換言すれば、部分集合生成部１１は、データの集合に属する全ての個別データから得られる一対の個別データの各組について、個別データ間の類似度を０に初期化する。

【００３０】

次に、部分集合生成部１１は、データの集合から部分集合を生成した回数を表す変数ｔを０に初期化する（ステップＡ２）。

【００３１】

続いて、部分集合生成部１１は、変数ｔと、Ｔとを比較し、ｔがＴ未満であるか否かを判定する（ステップＡ３）。変数ｔがＴ以上であるならば（ステップＡ３におけるＮｏ）、処理を終了する。また、変数ｔがＴ未満であるならば（ステップＡ３におけるＹｅｓ）、部分集合生成部１１は、データの集合からＭ個の個別データを選択することによって部分集合を生成する（ステップＡ４）。すなわち、部分集合の要素となるＭ個の個別データをデータの集合から選択すればよい。ステップＡ４の処理はｔの値がインクリメントされる毎に行われるが、２回目以降の部分集合の生成時において、新たに生成する部分集合に属する個別データと、既に生成された部分集合に属する個別データとが重複していてもよい。このように、部分集合生成部１１は、部分集合同士での個別データの重複を許して部分集合を生成する。

【００３２】

ステップＡ４において、データの集合から個別データを選択する際に、例えば、個別データをランダムサンプリングしてもよい。すなわち、部分集合生成部１１は、データの集合から１個の個別データをランダムに選択することをＭ回繰り返すことによって、Ｍ個の個別データを選択してもよい。

【００３３】

ただし、Ｍ個の個別データの選択方法はランダムサンプリングに限定されない。例えば

、部分集合生成部１１は、最初にステップＡ４を実行するときには、１番目からM番目までの個別データを選択し、次にステップＡ４を実行するときには、Ｍ＋１番目から２・Ｍ番目までの個別データを選択するというように、個別データに予め定められた順番に従って、ステップＡ４毎にM個ずつ個別データを選択してもよい。

【００３４】

なお、部分集合に含める個別データは、データの集合に属する個別データに含まれる属性のうち一部の属性と、クラスを抽出したデータであってもよい。例えば、図２に示す例では、個別データには「天気」、「天候」、「湿度」、「風が強いか否か」という４つの属性が含まれているが、部分集合に含める個別データを選択するときには、この４つの個別データのうちの一部（例えば、「天気」、「天候」のみ）と、クラスとを抽出し、その属性およびクラスからなるデータを部分集合の要素としてもよい。データの集合に属する個別データに含まれる属性がＡ個であるとき、部分集合に属する個別データに含まれる属性をいくつにするか、また、どの属性を部分集合に属する個別データに含めるかについては、例えば、キーボード等の入力装置（図示せず）を介して、データ類似度計算システムのユーザによって入力されてもよいし、他の態様で指定されてもよい。

【００３５】

ステップＡ４の後、分類器生成部１２は、直前のステップＡ４で生成された部分集合に基づいて分類器を生成する（ステップＡ５）。分類器として、例えば決定器やサポートベクタマシンを生成すればよい。どのような分類器を用いるかについては、キーボード等の入力装置（図示せず）を介して、データ類似度計算システムのユーザによって入力されてもよいし、他の態様で指定されてもよい。

【００３６】

以下、分類器として決定木を用いる場合を例にして、分類器の具体例を説明する。本例では、「購入する、購入しない」をクラスとし、「天気」、「気温」、「湿度」、「風が強いか否か」を属性とする場合を例にして説明する。また、ここでは、図２に示す１４個の個別データからなる部分集合が生成されている場合を例にする。図４は、分類器の例を示す説明図であり、ここでは、決定木を分類器としている。図２に示すような既知の属性およびクラスの組み合わせがあれば、その既知の属性およびクラスから決定木を生成することができる。以下の説明において、属性の値（すなわち属性値）に応じて、決定木におけるノードを枝分かれさせることを分割という。

【００３７】

図２に示す例では、各行が、「購入する」または「購入しない」というクラスが付された個別データに相当する。また、説明を簡単にするため、「購入する」というクラスを正（＋）のクラスと表し、「購入しない」というクラスを負（－）のクラスと表す場合がある。決定木では、クラス毎（例えば、「購入する（正）」、「購入しない（負）」というクラス毎）に個別データ数をまとめた情報をノードとする。例えば、図２に示すルートのノードでは、「する：９，しない：５」という情報をノードとしている。

【００３８】

分類器生成部１２は、ステップＡ４で生成された部分集合が与えられると、どの属性で最初にルートのノードを分割させるかを決定する。また、個別データに含まれている属性の個数をＲ個とした場合、各属性を属性１～Ｒとする。このとき、分類器生成部１２は、属性１～Ｒの各属性について、分割時の評価値を計算し、その評価値が最大の属性を、分割に最も適した属性として選択する。ここでは、分割前のノードのエントロピーと、分割後のエントロピーの差を評価値とする場合を例にするが、他の計算方法で評価値を求めてもよい。ノードのエントロピーは、クラスが正（＋）の個別データの割合をｑとし、クラスが負（－）の個別データの割合を１－ｑとすると、－ｑｌｏｇｑ－（１－ｑ）ｌｏｇ（１－ｑ）で表される。分割後のノードのエントロピーは、分割後の各ノードのエントロピーの加重平均である。

【００３９】

例えば、ステップＡ４で生成された部分集合において、「正」が９データあり、「負」

が５データあるとするとルートのノードは「正：９，負：５」となる。この場合、正（＋）のデータが９データあり、負（－）のデータが５データあるので、ルートのノードのエントロピーは、－（９／１４）×ｌｏｇ（９／１４）－（５／１４）×ｌｏｇ（５／１４）＝０．９４０となる。ただし、本例ではｌｏｇの底を２とする。

【００４０】

分類器生成部１２は、一つの属性でルートのノードを分割して得られるノードを求める。すなわち、その属性の属性値毎に、正および負の個別データ数を表す情報（ノード）を生成する。例えば、その属性１のとり得る値が「０」または「１」であり、属性１の値が「０」のときには、正が５データあり、負が２データあるとし、属性１の値が「１」のときには、正が０データあり、負が７データあるとする。この場合、属性１の値が「０」か「１」かで分岐するノードとして、「正：５，負：２」というノードと、「正：０，負：７」というノードとを生成する。分類器生成部１２は、分割後の各ノードのエントロピーを計算し、分割後の各ノードにおける正または負としてカウントされる個別データ数に応じて各ノードのエントロピーの加重平均を求める。上記の例では「正：５，負：２」というノードにおいても、「正：０，負：７」というノードにおいても個別データの総数は７であるので、加重平均を行う際の重み付け係数は各ノードでいずれも（７／１４）となる。従って、本例の場合、分類器生成部１２は、分割後のエントロピーを以下のように計算する。

【００４１】

（７／１４）×｛－（５／７）×ｌｏｇ（５／７）－（２／７）×ｌｏｇ（２／７）｝
＋（７／１４）×｛－（０／７）×ｌｏｇ（０／７）－（７／７）×ｌｏｇ（７／７）｝
＝０．４３２

【００４２】

ただし、ｌｏｇの係数が０となる場合、その項の値は０とする。上記の例では、－（０／７）×ｌｏｇ（０／７）の値は０としている。

【００４３】

従って、本例の場合、分類器生成部１２は、属性１で分割した場合の評価値を、０．９４０－０．４３２＝０．５０８と計算する。

【００４４】

分類器生成部１２は、属性１だけでなく、他の属性についても同様に、その属性で分割したときの評価値を計算し、評価値が最大となる属性で分割すると決定する。このようにして、ルートのノードを分割する属性を決定する。

【００４５】

なお、上記の属性１の例では、属性１のとり得る値が「０」または「１」の二つだけである場合を示した。属性値が年齢であり、その値が２０，２１，２２のように連続する値の場合には、どの属性値で分割させるのかも決める。この場合、分類器生成部１２は、各属性値間の中間値をしきい値とし、各しきい値毎に、その「しきい値以下」および「そのしきい値より大」とに分割させた場合の評価値を求める。そして、評価値が最大となる場合を選択することによって、どの属性値で分割させるのかも決定する。例えば、属性値が２０，２１，２２，・・・と連続する場合では、「２０．５以下」および「２０．５より大」で分割した場合の評価値、「２１．５以下」および「２１．５より大」で分割した場合の評価値等をそれぞれ計算し、評価値が最も高くなるように分割すればよい。

【００４６】

分類器生成部１２は、分割後の各ノードについても、上記と同様の処理を行い、次にどの属性で分割するのかを決定する処理を順次、繰り返す。また、分類器生成部１２は、所定の条件が満たされたときには、ノードの分割を停止する。所定の条件とは、例えば、「ノードにおける個別データのクラスが全て同じになる」という条件や、「ノードにおける正または負としてカウントされる個別データ数が所定数（例えば２）以下になる」という条件を用いてよい。前者の条件を採用すると、ノードにおける個別データが全て正または負になると、そのノードの分割を継続しない。このように、分類器生成部１２は、ルート

のノードから順次、分割を繰り返し、木構造の決定木を生成する。

【００４７】

また、分類器生成部１２は、上記のように、木構造の決定木を生成した後、その決定木に対する枝刈りを行う。決定木において、分割されて生成された最終的なノードを葉と呼ぶ。ある葉に分類されたデータ数がＤであるとする（すなわち、正または負としてカウントされる個別データ数がＤであるとする）。この葉に分類されたＤデータ中、Ｅデータが誤りであるとする。この仮定では、Ｄ回の試行中、誤りという事象をＥ回観測したとみなし、大きさＤの標本で、誤りという事象が起きる確率がｒである二項分布と考えることができる。予め与えられた信頼度ＣＦに対して、ｒの上限をＵ＿ＣＦ（Ｅ，Ｄ）と表すことにすると、Ｄデータでの誤りの発生する期待値は、Ｄ×Ｕ＿ＣＦ（Ｅ，Ｄ）となる。分類器生成部１２は、子のノードが全て葉である親のノードに対し、親における誤りの期待値（誤りの発生する期待値）と、子である葉の誤りの期待値の合計とを比較する。そして、子での期待値の合計の方が親の誤りの期待値よりも大きければ、分類器生成部１２は、葉を縮退して、その親を葉とする。分類器生成部１２は、この処理を順次繰り返すことで、決定木全体の葉の枝刈りを行う。

【００４８】

葉を縮退する場合、分類器生成部１２は、葉を削除して、その削除した葉の親のノードを葉とすればよい。例えば、図４に例示する決定木において、「湿度」という属性の値に応じて分割したノードを縮退する場合、分類器生成部１２は、湿度の属性値が７０％以下となっている個別データ数を表すノード「する：２，しない：０」と、湿度の属性値が７０％より高くなっている個別データ数を表すノード「する：０，しない：３」とを削除して、その２つのノードの親ノード「する：２，しない：３」を葉とすればよい。

【００４９】

分類器として決定木を生成する場合、例えば、上記のように、決定木を定めて枝刈りを行うことで、決定木を生成すればよい。

【００５０】

分類器生成部１２は、ステップＡ４で生成された部分集合に属するデータのうち、特定のクラスのデータを重く加重してから決定木を生成してもよい。ここで、加重とは、データに対する重みを設定することである。例えば、部分集合に属するデータのうち、予め定められた特定のクラスの個別データの数が数倍になるように、そのクラスの個別データの複製を作成することで、加重を行ってもよい。なお、特定のクラスの個別データ数を何倍にするかは予め定めておけばよい。そして、特定のクラスの個別データ数を増やすように、そのクラスの個別データを複製した後に、分類器を生成してもよい。加重の対象とするクラスおよび加重量（例えば、データを何倍に増やすか等）は、キーボード等の入力装置（図示せず）を介してデータ類似度計算システムのユーザにより入力されてもよいし、他の様態で指定されてもよい。

【００５１】

ステップＡ５で分類器が生成されると、自己評価部１３は、その分類器を用いて、ステップＡ３で生成された部分集合に属する個々のデータのクラスを判定する（ステップＡ６）。ここでも、分類器が決定木である場合を例にして説明する。分類器となる決定木を生成した場合、その決定木と、既知の属性（部分集合に属する個別データの属性）とから、クラスを予測する。このとき、自己評価部１３は、決定木のルートのノードを起点として、ノードを分割する際に用いた属性に関して個別データの属性値を参照し、その属性値に応じて子ノードを辿る。自己評価部１３は、子ノードを辿っていき、葉のノードまで辿ったならば、葉のノードでカウント数の多い方のクラスを、分類器を用いた判定の結果とすればよい。

【００５２】

図５は、分類器から判定されたクラスの例を示す説明図である。図５に示す各属性の属性値は、図２に示す各属性値と同一である。また、図５では、図４に例示する決定木（分類器）を用いて、その属性値からクラスを判定した結果を示している。例えば、個別デー

タにおける属性「天気」の属性値が「雨」であるとする。すると、ルートのノードから、「する：３、しない：２」という子ノードを辿る（図４参照）。そのノードが葉であるので、自己評価部１３は、「する：３、しない：２」というカウント数により、「購入しない」というクラスであると判定する。このように判定した結果、図５に示す６番目および最後の個別データのクラスの判定結果は、図２に示す元のクラスと異なっている（図２、図５参照）。

【００５３】

ステップＡ６で部分集合に属する各個別データのクラスが判定された後、類似度算出部１４は、データの集合に属する個別データを順番に指定するための第１の変数ｉを１に初期化する（ステップＡ７）。続いて、類似度算出部１４は、変数ｉの値が個別データの総数Ｎ以下であるか否かを判定する（ステップＡ８）。変数ｉの値がＮを超えていれば（ステップＡ８におけるＮｏ）、変数ｔを１インクリメントし（ステップＡ８）、その後、データ類似度計算システムはステップＡ３以降の処理を再度繰り返す。

【００５４】

変数ｉの値がＮ以下であるならば（ステップＡ８におけるＹｅｓ）、類似度算出部１４は、変数ｉによって定まるｉ番目の個別データが、直近のステップＡ４で生成された部分集合に含まれているか否かを判定する（ステップＡ１０）。ｉ番目のデータが部分集合に含まれていないならば（ステップＡ１０におけるＮｏ）、類似度算出部１４は、変数ｉを１インクリメントし（ステップＡ１１）、ステップＡ８以降の処理を再度繰り返す。

【００５５】

ｉ番目のデータが部分集合に含まれているならば（ステップＡ１０におけるＹｅｓ）、データの集合に属する個別データを順番に指定するための第２の変数ｊの値をｉ＋１に設定する（ステップＡ１２）。すなわち、ｊにｉ＋１を代入する。

【００５６】

ステップＡ１２の後、類似度算出部１４は、変数ｊの値がＮ以下であるか否かを判定する（ステップＡ１３）。変数ｊの値がＮを超えていれば（ステップＡ１３におけるＮｏ）、類似度算出部１４は、変数ｉを１インクリメントし（ステップＡ１１）、ステップＡ８以降の処理を再度繰り返す。

【００５７】

変数ｊの値がＮ以下であるならば（ステップＡ１３におけるＹｅｓ）、類似度算出部１４は、変数ｊによって定まるｊ番目の個別データが、直近のステップＡ４で生成された部分集合に含まれているか否かを判定する（ステップＡ１４）。ｊ番目の個別データが部分集合に含まれていないならば（ステップＡ１４におけるＮｏ）、類似度算出部１４は、変数ｊを１インクリメントし（ステップＡ１７）、ステップＡ１３以降の処理を繰り返す。

【００５８】

ｊ番目の個別データが部分集合に含まれているならば（ステップＡ１４におけるＹｅｓ）、類似度算出部１４は、ｉ番目の個別データの属性から判定されたクラス（$PredC_i$と記す）と、ｊ番目の個別データの属性から判定されたクラス（$PredC_j$と記す）とをとを比較し、両者が同じクラスであるか否かを判定する（ステップＡ１５）。$PredC_i$および$PredC_j$は、直近のステップＡ６で、ｉ番目およびｊ番目の個別データについて、分類器を用いて属性から判定されたクラスである。$PredC_i$と$PredC_j$とが同じクラスであるということは、ｉ番目の個別データおよびｊ番目の個別データが、同一のクラスと判定されたデータ同士であるということを意味する。

【００５９】

$PredC_i$と$PredC_j$とが異なっていれば（ステップＡ１５におけるＮｏ）、類似度算出部１４は、変数ｊを１インクリメントし（ステップＡ１７）、ステップＡ１３以降の処理を繰り返す。

【００６０】

$PredC_i$と$PredC_j$とが同じクラスであれば（ステップＡ１５におけるＹｅｓ）、ｉ番目の個別データとｊ番目の個別データの類似度Ｓｉｍ（ｉ，ｊ）に所定数を加算

する（ステップＡ１６）。本例では、この所定数を１とし、ステップＡ１６でＳｉｍ（ｉ，ｊ）に１を加算する場合を例にする。ステップＡ１６の後、類似度算出部１４は、変数ｊを１インクリメントし（ステップＡ１７）、ステップＡ１３以降の処理を繰り返す。

【００６１】

上記のように、１つの部分集合が生成されたとき、その部分集合に属している１対の個別データであって、$PredC_i$と$PredC_j$とが同じクラスとなっている個別データ間の類似度Ｓｉｍ（ｉ，ｊ）の値を１増加させていく。また、一方あるいは両方が部分集合に属していない個別データ同士の場合には、ステップＡ１６に移行しないので、その個別データ間の類似度は増加しない。また、部分集合に属している１対の個別データであっても、属性から判定されたクラスが異なる場合にも、ステップＡ１６に移行せず、その個別データ間の類似度は増加しない。

【００６２】

そして、データ類似度計算システムは、変数ｔをインクリメントして（ステップＡ８）、部分集合を生成すると（ステップＡ４）、ステップＡ５以降の処理を実行する。このとき、各個別データの組について一律にステップＡ１６の処理を実行するわけではない。個別データの重複を許して部分集合の生成を複数回生成したときに、ステップＡ１６が行われた回数の多い個別データの組については、Ｓｉｍ（ｉ，ｊ）の値も大きくなる。一方、ステップＡ１６が行われた回数の少ない個別データの組については、Ｓｉｍ（ｉ，ｊ）の初期値からの増加量は少ない。部分集合の生成回数ｔが上限値Ｔに達したときにおける各データの対毎のＳｉｍ（ｉ，ｊ）が、対をなす個別データ間の類似度となる。

【００６３】

また、上記の第１の実施の形態では、二つのデータがいずれも部分集合に属し、その二つのデータに対して判定されたクラスが同一であるという条件が満たされたときに、その二つのデータ類似度Ｓｉｍ（ｉ，ｊ）に対して所定値を加算する。この条件だけでなく、さらに、二つのデータに対して判定されたクラスが特定のクラス（例えば、「購入する」）であるという条件を満たしている場合に、二つのデータ類似度Ｓｉｍ（ｉ，ｊ）に対して所定値を加算し、他の場合には、Ｓｉｍ（ｉ，ｊ）への加算を行わなくてもよい。この場合、二つのデータに対するクラスの判定結果が同一であっても、そのクラスが特定のクラス出ない場合、類似度に対する加算を行わない。

【００６４】

上記のように類似度を求める場合、類似度算出部１４は、ステップＡ１５において、$PredC_i$と$PredC_j$とが同一であり、かつ、$PredC_i$および$PredC_j$が特定のクラス（例えば、「購入する」）であるか否かを判定すればよい。そして、この条件を満たしている場合に、ステップＡ１６を行い、満たしていない場合には、ステップＡ１７に移行すればよい。

【００６５】

次に、第１の実施形態の効果について説明する。

第１の実施形態によれば、与えられたデータの集合に属するデータ同士の組に対して、それぞれ類似度の初期値を定める。そして、そのデータの集合から部分集合を生成し、その部分集合に含まれる各データに定められている属性（例えば、顧客の特徴や販売条件）およびクラスから、分類器を生成する。さらに、分類器を用いて、その属性からクラスを判定し、部分集合に含まれているデータ同士であって、判定されたクラスが同じデータの類似度に所定値を加算する。上記の部分集合生成以後の処理を複数回繰り返すことで、各データ同士の類似度を決定する。本願発明では、このような処理によって、自動的に、各データ間の類似度を算出することができる。

【００６６】

また、本実施形態において、部分集合生成時（ステップＡ４）では、個別データを予め定められた数Ｍだけ選択すればよく、既に生成された部分集合に属している個別データを重複を許して選択する。例えば、ランダムサンプリングを行ってもよい。このように、Ｍ個の個別データを選択することを複数回行って、類似度を生成する。また、類似度が求ま

れば、類似度に基づいて個別データを分類することができる。特許文献１に記載されたシステムのように、分析者が手動で顧客の分類を行う場合には、顧客の分類に対する特別なスキルが必要となったり、試行錯誤して分析を行って手間がかかったりすることがあるが、上記のように類似度を求めれば、システムのユーザに特別なスキルがなくても、ユーザの手間をかけずに、分類に用いるための類似度を算出することができる。

【００６７】

また、各データが個々の顧客に対応し、データの属性が顧客の特徴や販売条件であり、データに「購入する」または「購入しない」というクラスが与えられている場合、商品またはサービスに対する顧客の行動（購入したか否か）に基づいて、データの類似度を求めることができる。また、データが顧客に対応しているので、顧客同士の類似度ということもできる。

【００６８】

以下に示す効果の説明では、データの属性が顧客の特徴であり、データと顧客が対応していて、データのクラスが顧客の購入行動（例えば「購入する」または「購入しない」）である場合を例にして説明する。

【００６９】

重複を許したランダムサンプリングにより部分集合を生成した場合、個々のデータ（本例では個々の顧客）は、部分集合にＭ／Ｎの確率で選択されることとなる。すなわち、Ｍ／Ｎの確率で部分集合に含められることになる。他のデータ（他の顧客）も、同様の確率で部分集合に含められる。データを組み合わせて得られる各データの組は、同じ確率で部分集合に属することとなる。従って、部分集合を作成する回数Ｔが十分大きく、仮に、元のクラスとステップＡ６で判定されたクラスが全て同一であれば、データ間の類似度はほぼ同じとなり、クラスが同一のデータを細かく分類することはできない。例えば、「購入する」というクラスのデータを細かく分類したり、「購入しない」というクラスのデータを細かく分類したりすることはできない。また、「購入する」というクラスのデータ（商品等を購入した顧客）と、「購入しない」というクラスのデータ（商品等を購入しなかった顧客）との類似度は０となる。これに対し、本発明の第１の実施形態では、部分集合に属する各個別データから分類器を生成し、その分類器を用いて、個別データの属性からクラスを判定することで、元のクラスと、判定後のクラスとがことなる場合を生じさせる機会を積極的に設けている。そして、判定後のクラスを用いて、個別データの類似度を求めることで、ｉ番目のデータとｊ番目のデータからなる各組に対して、それぞれ類似度Ｓｉｍ（ｉ，ｊ）を定めることができる。

【００７０】

また、本発明では、互いに関連のある属性が多く存在しても、それらの関連ある属性の影響を強く受けることなく類似度を算出できる。例えば、図１２に示すデータ番号１～１０の１０個のデータが与えられているとする。図１２に示すデータ番号｛１，２，７，９，１０｝のデータを部分集合とすると、例えば、分類器として「属性ａの値が１ならばクラスは『購入する』であり、属性ａの値が２ならばクラスは『購入しない』である」という決定木が生成され、データ番号｛１，２｝のデータ間の類似度や、データ番号｛７，９，１０｝の各データ間の類似度が加算される。また、例えば、データ番号｛１，２，６，７，９，１０｝のデータを部分集合とする。この場合、例えば、分類器として「属性ｂの値が１ならばクラスは『購入する』であり、属性ｂの値が２ならばクラスは『購入しない』である」という決定木が生成され、データ番号｛１，２，６｝の各データ間の類似度や、データ番号｛７，９，１０｝の各データ間の類似度が加算される。

【００７１】

一般に、決定木には複数の属性が現れるが、属性に互いに関連性がある場合、決定木に現れる属性の数は減る。上記のデータ番号｛１，２，６，７，９，１０｝を部分集合とする場合を例にすると、属性ｂを用いて個別データを２つに分類した場合と、属性ｃ、ｄ、あるいはｅを用いて個別データを２つに分類した場合とで、個別データの分類のされ方が全く同一になる。よって、属性ｂで分割した後、属性ｃ、ｄ、あるいはｅで分割すること

はない。具体例を挙げると、データ番号｛１，２，６，７，９，１０｝のデータを属性ｂで分割すると、属性ｂの値が１のノードにはデータ番号｛１，２，６｝のデータが含まれ、属性ｂの値が２のノードにはデータ番号｛７，９，１０｝が含まれる。データ番号｛１，２，６｝のノードをｃで分割する場合、そのデータ番号｛１，２，６｝のデータは全て属性ｃの値が１となり、属性ｃの値が２となるデータはない。属性ｂの値が２のノードでも、属性ｃの値が一方に偏っている。属性ｂでの分割後、属性ｄやｅで分割する場合も同様である。

【００７２】

このように、属性に互いに関連性がある場合、決定木に現れる属性の数が減り、関連のある属性の影響を強く受けることがなくなる。上記のデータ番号｛１，２，６，７，９，１０｝を部分集合とする例では、例えば、決定木には属性ｂが現れるが、属性ｂと同時に他の属性ｃ、ｄ、ｅは決定木に現れず、属性ｂに関連する属性ｃ、ｄ、ｅに影響されることがない。そのため、データに関連ある属性が多く存在していても、それらの関連のある属性の影響を強く受けることなく、類似度を求めることができる。

【００７３】

また、二つのデータがいずれも部分集合に属し、その二つのデータに対して判定されたクラスが同一であるという条件だけでなく、さらに、その二つのデータに対して判定されたクラスが特定のクラス（例えば、「購入する」）であるという条件を満たしているときにのみ、その二つのデータの類似度Ｓｉｍ（ｉ，ｊ）への加算（ステップＡ１６）を行う場合、以下の効果が得られる。すなわち、特定のクラスのデータの持つ特徴の違いを重視して類似度を算出できる。ここでも、データと顧客とが対応している場合を例にして説明する。ステップＡ１６に移行する条件として、二つのデータに対して判定されたクラスが特定のクラスであるという条件も加えると、特定のクラスの顧客が持つ特徴の違いを重視して類似度を算出できる。例えば、商品またはサービスを購入した顧客が複数の特徴のいずれかを持っていて、まだその商品またはサービスを購入していないが、既に購入済みの顧客と特徴が近い顧客（購入の見込みのある顧客）が複数の特徴のうちのいずれかを持っていそうな場合がある。そのような場合において、購入していない顧客との類似度は重視せず、購入した顧客との類似度を重視したいと分析者が考える場合がある。そのような場合、ステップＡ６で「購入する」という特定のクラスであると判定された顧客のデータと、ステップＡ６で「購入する」と判定された「購入の見込みのある顧客」のデータとの類似度について加算して、「購入する」というクラスの顧客が持つ特徴の違いを重視して類似度を算出することができる。

【００７４】

また、分類器生成部１２は、分類器を作成するときに、ステップＡ４で生成された部分集合に属するデータのうち、予め与えられたクラスが特定のクラスとなっているデータに対して加重を行ってから、分類器を作成してもよい。例えば、部分集合に属するデータのうち、予め与えられたクラスが「購入する」となっているクラスのデータのデータ数が、指定された分だけ増えるように、そのデータの複製を作成して、その複製したデータも用いて分類器を生成してもよい。そのような分類器を用いて類似度を算出すれば、その類似度に基づいてデータを分類する場合、特定のクラス以外のクラスのデータが持つ特徴の違いを重視してクラスタ（グループ）に分類することができる。例えば、「購入する」というクラスの顧客（データ）と、「購入しない」というクラスの顧客の数が同じくらいであるとする。データに加重を行わない場合には、いずれのクラスのデータも混ざったクラスタが生成されがちになる。特定のクラスのデータを重く加重すると、特定のクラスのデータの間の類似度がどの組み合わせでも高くなることで１つのクラスタとなる。一方、特定のクラス以外のクラスのデータは、そのように類似度が高くなることはなく、１つのクラスタにまとめられずに、複数のクラスタに分類することができる。なお、類似度を用いてデータをクラスタ（グループ）に分類する処理については、第２の実施形態で説明する。

【００７５】

実施形態２．

図６は、本発明の第２の実施形態の例を示すブロック図である。第２の実施形態のデータ類似度計算システム２０は、部分集合生成部１１と、分類器生成部１２と、自己評価部１３と、類似度算出部１４と、類似度クラスタリング部２１とを備える。第１の実施形態と同様の構成要素については、図１と同一の符号を付し、詳細な説明を省略する。第２の実施形態のデータ類似度計算システム２０は、第１の実施形態と同様に、データの集合に属する個別データ間の類似度を計算した後、その類似度を用いて個別データを分類する。従って、第２の実施形態のデータ類似度計算システム２０は、分類システムと称することができる。以下、本実施形態において、個別データのグループをクラスタと記す。

【００７６】

類似度クラスタリング部２１は、類似度算出部１４が求めた個別データ間の類似度に基づいて、データの集合に属する個別データを複数のクラスタに分類する。類似度クラスタリング部２１には、目標とするクラスタ数が入力され、類似度クラスタリング部２１は、そのクラスタ数になるように個別データを分類する。目標とするクラスタ数は、データ類似度計算システム２０に設けられるキーボード等の入力装置（図示せず）を介して、データ類似度計算システムのユーザによって入力されてもよい。あるいは、他の態様でクラスタ数が指定されてもよい。

【００７７】

部分集合生成部１１、分類器生成部１２、自己評価部１３、類似度算出部１４および類似度クラスタリング部２１は、例えば、プログラム（データ類似度計算プログラム）に従って動作するＣＰＵによって実現される。その場合、ＣＰＵがプログラムに従って、部分集合生成部１１、分類器生成部１２、自己評価部１３、類似度算出部１４および類似度クラスタリング部２１として動作すればよい。

【００７８】

次に、第２の実施形態の動作について説明する。

データ集合に属する各個別データ間の類似度を求める処理は、第１の実施形態と同様である。部分集合生成部１１、分類器生成部１２、自己評価部１３および類似度算出部１４がそれぞれ第１の実施形態と同様に動作し、例えば、図３に示す処理を行って、ｉ，ｊの組毎に、個別データ間の類似度Ｓｉｍ（ｉ，ｊ）を求めればよい。

【００７９】

それぞれの個別データ間の類似度Ｓｉｍ（ｉ，ｊ）が算出された後、類似度クラスタリング部２１は、個別データの各組における類似度を用いて、個別データをクラスタリングする（すなわち分類する）。

【００８０】

類似度が与えられたときのクラスタリング方法には、様々な方法がある。例えば、階層的クラスタリング法として最短距離法、最長距離法、郡平均法、ウォード法等があり、非階層的クラスタリング法としてＫ平均法等がある。類似度クラスタリング部２１は、いずれの方法で個別データを分類してもよい。また、どの方法でクラスタリングを行うかの指定が、キーボード等の入力装置を介してデータ類似度計算システムのユーザにより入力されてもよい。あるいは、他の態様で指定されてもよい。

【００８１】

以下、最短距離法によって個別データをクラスタリングする場合を例にして、個別データのクラスタリング処理を説明する。図７は、最短距離法によって個別データをクラスタリングする処理経過の例を示すフローチャートである。なお、以下の説明では、目標とするクラスタ数をＫ個とする。目標とするクラスタ数Ｋは、例えば、キーボード等の入力装置を介して、予め類似度クラスタリング部２１に入力される。

【００８２】

類似度クラスタリング部２１は、データの集合に属する個別データをそれぞれ１個だけ含むクラスタを、各個別データ毎に定める（ステップＢ１）。従って、類似度クラスタリング部２１は、ステップＢ１において、データの集合に属する個別データの総数Ｎと等しいＮ個のクラスタを定めることになる。このＮ個のクラスタに属する個別データは、クラ

スタ毎に異なっている。ステップＢ１の後、ステップＢ２に移行する。

【００８３】

ステップＢ２では、類似度クラスタリング部２１は、二つのクラスタからなるクラスタの各組についてそれぞれ、クラスタ間の類似度を求め、最もクラスタ間の類似度が高い二つのクラスタを特定し、その二つのクラスタを一つのクラスタに併合する（ステップＢ２）。ステップＢ２に移行した時点でのクラスタ数をＬとすると、Ｌ個のクラスタから二つのクラスタを取り出す${}_{L}C_{2}$個の組毎に、クラスタ間の類似度を求め、最もクラスタ間の類似度が高い二つのクラスタを一つのクラスタに併合すればよい。なお、最初にステップＢ１からステップＢ２に移行したときには、Ｌ＝Ｎである。

【００８４】

また、二つのクラスタからなる組において、その二つのクラスタ間の類似度を定める方法の例を以下に示す。類似度を求める対象となる二つのクラスタの一方をC_1と記し、他方をC_2と記す。また、その二つのクラスタC_1，C_2間の類似度をＳ（C_1，C_2）と記す。類似度クラスタリング部２１は、例えば、C_1に属する個別データと、C_2に属する個別データとの各組み合わせにおける個別データ間の類似度のうち、最大値をC_1，C_2間の類似度をＳ（C_1，C_2）と定めればよい。すなわち、C_1から取り出した一つの個別データをx_1とし、C_2から取り出した一つの個別データをx_2とし、x_1，x_2の類似度をＳ（x_1，x_2）とすると、類似度クラスタリング部２１は、式（１）に示すようにＳ（x_1，x_2）の最大値をＳ（C_1，C_2）と定めればよい。

【００８５】

【数１】

$$S(C_1, C_2) = \max_{x_1 \in C_1,\ x_2 \in C_2} S(x_1, x_2) \qquad \text{式（１）}$$

【００８６】

また、二つのクラスタを一つのクラスタに併合するとは、二つのクラスタに属する各個別データを一つのクラスタにまとめることである。二つのクラスタを一つのクラスタに併合することにより、クラスタの総数が１つ減少する。

【００８７】

ステップＢ２で二つのクラスタを一つのクラスタに併合した後、類似度クラスタリング部２１は、クラスタ数が目標数Ｋになったか否かを判定する（ステップＢ３）。クラスタ数が目標数Ｋまで減っていなければ（ステップＢ３におけるＮｏ）、ステップＢ２以降の処理を繰り返す。クラスタ数が目標数Ｋとなっていれば（ステップＢ３におけるＹｅｓ）、Ｋ個のクラスタが得られ、個別データがＫ個に分類されているので、処理を終了する。

【００８８】

第２の実施形態によれば、データの集合に含まれる個別データを目標数のクラスタに分類することができる。

【００８９】

例えば、データの属性が顧客の特徴であり、データと顧客が対応していて、データのクラスが顧客の購入行動（「購入する」または「購入しない」等）を表している場合、商品またはサービスに対する顧客の購入行動に応じて、自動的に顧客を分類することができる。

【００９０】

また、図７に例示するように、最初に各個別データをそれぞれ別々のクラスタに振り分け、クラスタの数が目標数となるまでクラスタを併合させていけば、目標数のクラスタに個別データを分類することができる。すなわち、特定のクラスタに個別データが集まってしまい個別データのクラスタ数が目標数に達しないということを防止することができる。

【００９１】

実施形態３．

図８は、本発明の第３の実施形態の例を示すブロック図である。第３の実施形態のデータ類似度計算システム３０は、部分集合生成部１１と、分類器生成部１２と、自己評価部１３と、類似度算出部１４と、類似度クラスタリング部２１と、属性データ分類部３１と、関連性算出部３２とを備える。第１の実施形態や第２の実施形態と同様の構成要素については、図１、図６と同一の符号を付し、詳細な説明を省略する。第３の実施形態のデータ類似度計算システム３０は、第２の実施形態と同様に類似度を用いて個別データを分類するので、分類システムと称することができる。なお、本実施の形態においても、類似度クラスタリング部２１によって分類された個別データのグループをクラスタと記す。また、第３の実施形態のデータ類似度計算システム３０は、類似度算出に対する属性の関連度を算出する。類似度算出に対する属性の関連度とは、類似度算出に対する属性の関連性の度合い（換言すれば、類似度算出に対して属性が影響を与える度合い）を示す数値である。

【００９２】

属性データ分類部３１には、属性およびその属性値に基づく分類方法を指定する情報（以下、分類方法指定情報と記す）が入力され、属性データ分類部３１は、指定された分類方法に従って、個別データを分類する。分類方法指定情報によって、属性データ分類部３１による個別データの分類数も定まる。分類方法指定情の例として、「属性名Ａの属性値がＴｈｒｅｓ以上である個別データをグループ１に分類し、属性名Ａの属性値がＴｈｒｅｓ未満である個別データをグループ２に分類する。」などの情報が挙げられる。この場合、個別データは、二つのグループ１，２に分類されることになる。分類方法指定情報は、例えば、キーボード等の入力装置（図示せず）を介して属性データ分類部３１に入力されるが、他の態様で入力されてもよい。

【００９３】

関連性算出部３２は、類似度クラスタリング部２１によって分類されたクラスタ（グループ）と、属性データ分類部３１によって分類されたグループとの関係に基づいて、類似度算出に対する属性の関連度（以下、単に属性の関連度と記す）を求める。この属性は、分類方法指定情報で指定された属性である。

【００９４】

部分集合生成部１１、分類器生成部１２、自己評価部１３、類似度算出部１４、類似度クラスタリング部２１、属性データ分類部３１および関連性算出部３２は、例えば、プログラム（データ類似度計算プログラム）に従って動作するＣＰＵによって実現される。その場合、ＣＰＵがプログラムに従って、部分集合生成部１１、分類器生成部１２、自己評価部１３、類似度算出部１４、類似度クラスタリング部２１、属性データ分類部３１および関連性算出部３２として動作する。

【００９５】

次に、第３の実施形態の動作について説明する。

類似度クラスタリング部２１が個別データを分類するまでの動作は、第２の実施形態と同様である。部分集合生成部１１、分類器生成部１２、自己評価部１３および類似度算出部１４は、第１および第２の実施形態と同様に動作し、例えば、図３に示す処理を行って、ｉ，ｊの組毎に、個別データ間の類似度Ｓｉｍ（ｉ，ｊ）を求めればよい。その後、類似度クラスタリング部２１は、第２の実施形態と同様に、指定されたクラスタ数に個別データを分類する。

【００９６】

図９は、類似度クラスタリング部２１による分類後に属性の関連度を求める動作の例を示すフローチャートである。ここでは、類似度クラスタリング部２１によってグループ分けされるクラスタ数と、属性データ分類部３１によってグループ分けされるグループ数が等しい場合を例に説明する。本例では、類似度クラスタリング部２１がデータの集合に属する個別データを二つのクラスタに分類するものとする。また、属性データ分類部３１には、例えば、「属性名Ａの属性値がＴｈｒｅｓ以上である個別データをグループ１に分類し、属性名Ａの属性値がＴｈｒｅｓ未満である個別データをグループ２に分類する。」と

いう分類方法指定情報が入力され、属性データ分類部３１が個別データを２つのグループに分類するものとする。

【００９７】

また、以下の説明において、変数ｋは、類似度クラスタリング部２１によってグループ分けされるクラスタを指定するための変数であり、本例では、ｋの取り得る値は１または２である。また、変数ｌは、属性データ分類部３１によってグループ分けされたグループを指定するための変数であり、本例では、ｌの取り得る値は１または２である。

【００９８】

属性データ分類部３１は、まず、ｋ＝１，２、ｌ＝１，２によって決まるｋとｌの組み合わせを順次定め、ｋ，ｌの組み合わせと一対一に対応する変数Ｎ［ｋ］［ｌ］を全て０に初期化する（ステップＣ１）。Ｎ［ｋ］［ｌ］は、ｋ番目のクラスタに属している個別データであって、ｌ番目のグループにも属している個別データの数を表す。

【００９９】

属性データ分類部３１は、個別データを指定するための変数ｉを１に初期化する（ステップＣ２）。

【０１００】

次に、属性データ分類部３１は、変数ｋの値を、個別データｉ（すなわち、ｉ番目の個別データ）が属するクラスタのクラスタ番号とする（ステップＣ３）。なお、例えば類似度クラスタリング部２１は、分類の結果得られたＫ個のクラスタに対して、クラスタを識別するための番号１～Ｋを割り当てる。個々のクラスタに割り当てられた番号がクラスタ番号である。ステップＣ３では、変数ｋに、データｉが属するクラスタのクラスタ番号を代入すればよい。本例ではクラスタ数は２であるので、クラスタ番号は１または２である。

【０１０１】

次に、属性データ分類部３１は、個別データｉの属性値（分類方法指定情報で指定された属性の属性値）に応じて、個別データｉをどのグループに含めるかを判定する（ステップＣ４）。すなわち、分類方法指定情報に従って、個別データｉに対する分類を行う。本例では、「属性名Ａの属性値がＴｈｒｅｓ以上である個別データをグループ１に分類し、属性名Ａの属性値がＴｈｒｅｓ未満である個別データをグループ２に分類する。」という分類方法指定情報に従って、個別データｉの属性Ａの属性値がＴｈｒｅｓ以上であるか否かを判定する。個別データｉの属性Ａの属性値がＴｈｒｅｓ以上であるならば（すなわち、個別データｉをグループ１に分類すると判定したならば）、変数ｌの値を１に設定する（ステップＣ５）。また、個別データｉの属性Ａの属性値がＴｈｒｅｓ未満であるならば（すなわち、個別データｉをグループ２に分類すると判定したならば）、変数ｌの値を２に設定する（ステップＣ６）。

【０１０２】

ステップＣ５またはステップＣ６の後、ステップＣ３で定められたｋと、ステップＣ５またはステップＣ６で定められたｌとの組み合わせに対応するＮ［ｋ］［ｌ］の値を１インクリメントする（ステップＣ７）。例えば、ステップＣ５からステップＣ７に移り、Ｎ［ｋ］［１］を１インクリメントした場合、ｋ番目のクラスタに属している個別データであって、１番目のグループに属している個別データを一つカウントして、そのカウント値を１増加させたことになる。また、ステップＣ６からステップＣ７に移り、Ｎ［ｋ］［２］を１インクリメントした場合、ｋ番目のクラスタに属している個別データであって、２番目のグループに属している個別データを一つカウントして、そのカウント値を１増加させたことになる。

【０１０３】

次に、属性データ分類部３１は、変数ｉの値を１インクリメントする（ステップＣ８）。そして、属性データ分類部３１は、変数ｉの値が個別データの総数Ｎ以下であるか否かを判定し（ステップＣ９）、ｉの値がＮ以下であれば（ステップＣ９におけるＹｅｓ）、ステップＣ３以降の処理を再度行う。従って、１番目からＮ番目までの各個別データに対

してステップＣ３以降の処理を行うことになり、ｋ番目のクラスタに属している個別データであって、ｌ番目のグループにも属している個別データの数Ｎ［ｋ］［ｌ］が、ｋ，ｌの組み合わせ毎に求められる。

【０１０４】

ｉの値がＮを超えていれば（ステップＣ９におけるＮｏ）、ステップＣ１０に移行する。

【０１０５】

ステップＣ１０において、関連性算出部３２は、Ｎ［ｋ］［ｌ］を用いて、類似度クラスタリング部による分類と属性データ分類部による分類との独立性を検定し、ｐ値を算出する（ステップＣ１０）。このｐ値を、分類方法指定情報で指定された属性の関連度とすることができる。

【０１０６】

関連性算出部３２は、例えば、類似度クラスタリング部２１が行った分類と属性データ分類部３１が行った分類とが独立であるという仮説によりｐ値を求める。この場合、関連性算出部３２は、まずｋ，ｌの各組み合わせに関して、Ｎ［ｋ］［ｌ］の期待値（Ｅ［ｋ］［ｌ］と記す）を計算する。上記の仮説のもとでは、関連性算出部３２は、以下に示す式（２）の計算によって各Ｎ［ｋ］［ｌ］の期待値Ｅ［ｋ］［ｌ］を求めればよい。

【０１０７】

【数２】

$$E[k][l]=\frac{\sum_{g=1}^{2}N[k][g]\times\sum_{h=1}^{2}N[h][l]}{N} \qquad \text{式（２）}$$

【０１０８】

そして、Ｎ［ｋ］［ｌ］およびＥ［ｋ］［ｌ］を用いて、以下に示す式（３）の計算によりχ$_0$^{2}を計算すると、χ$_0$^{2}はχ^{2}分布に従う。

【０１０９】

【数３】

$$\chi_0^2=\sum_{k=1}^{2}\sum_{l=1}^{2}\frac{(N[k][l]-E[k][l])^2}{E[k][l]} \qquad \text{式（３）}$$

【０１１０】

そこで、関連性算出部３２は、式（３）の計算によりχ$_0$^{2}を求め、χ^{2}分布表を参照してｐ値を決定すればよい。このｐ値が、分類方法指定情報で指定された属性の関連度である。なお、χ^{2}分布表は、例えば、予めデータ類似度計算システム３０が備える記憶装置（図示せず）に記憶させておけばよい。

【０１１１】

χ^{2}分布表からｐ値を決定する際に用いる自由度は、類似度クラスタリング部２１によって分類されるクラスタ数－１である。従って、本例では、クラスタ数＝２であるので、自由度は、２－１＝１とすればよい。

【０１１２】

図１０は、χ^{2}分布表の例を示す説明図である。図１０では、自由度をνで表し、ν＝１，２，３の場合を例示しているが、ν＝４以上の場合も含めておく。図１０に示すχ^{2}分布表の最上段の値「０．７」、「０．５」、「０．３」、「０．２」、「０．１」等はｐ値である。例えば、式（３）でχ$_0$^{2}を計算した結果、関連性算出部３２は、χ$_0$^{2}

１．０７であったとする。本例では、χ_0^2は、自由度１のχ^2分布に従うので、自由度１におけるχ_0^2　１．０７に応じたｐ値“０．３”をχ^2分布表から特定し、そのｐ値“０．３”を、類似度算出に対する属性Ａの関連度とすればよい。

【０１１３】

以上のように、属性Ａの関連度を求めることができる。他の属性の関連度も、分類方法指定情報を入力して求めることができる。

【０１１４】

また、関連性算出部３２は、例えば、χ_0^2を計算した後に以下に示す式（４）の計算を行って、クラメールの関連係数を求め、そのクラメールの関連係数を、類似度算出に対する属性の関連度としてもよい。

【０１１５】

【数４】

$$\sqrt{\frac{\chi_0^2}{N}} \qquad \text{式（４）}$$

【０１１６】

クラメールの連関係数は、０から１までの数値であり、１に近いほど強く関連していることを示す。

【０１１７】

また、関連性算出部３２が類似度算出に対する属性の関連度を求めた後、その属性の関連度を、ディスプレイ装置またはプリンタ装置等の出力装置（図示せず）に出力させてから、終了してもよい。ユーザが入力装置（図示せず）を介して、属性を指定した分類方法指定情報を入力し、データ類似度計算システム３０がその属性の関連を求めてもよい。そして、関連性算出部３２は、そのようにして求めた各属性の関連度をディスプレイ装置またはプリンタ装置等の出力装置（図示せず）に出力させる際、表形式やグラフ形式で出力させてもよい。

【０１１８】

次に、第３の実施形態の効果について説明する。

本発明によれば、類似度によるクラスタリングと属性値との関連の度合いを調べることができる。例えば、商品またはサービスに対する顧客の行動に応じて算出した類似度によるクラスタリングと属性値との関連を調べることができる。具体例を挙げると、例えば、類似度によるクラスタリングと男女（性別）との関連性が高いかどうかを調べることができ、商品またはサービスに対する顧客の行動に応じて、顧客を分類する際に関連のある顧客の特徴や販売条件を抽出することができる。これにより、商品またはサービスの今後ターゲットとすべき顧客等を分析することができる。

【０１１９】

次に、本発明の概要について説明する。図１１は、本発明の概要を示すブロック図である。本発明のデータ類似度計算システム８０は、部分集合生成手段８１と、分類器生成手段８２と、クラス判定手段８３と、類似度算出手段８４とを備える。

【０１２０】

部分集合生成手段８１（例えば、部分集合生成部１１）は、データの特徴を示す属性とデータの類別を示すクラスとを含むデータの集合から、データの重複を許してその集合の部分集合を複数回生成する。

【０１２１】

分類器生成手段８２（例えば、分類器生成部１２）は、部分集合が生成される毎に、属性からクラスを判定するルールである分類器（例えば、決定木）を、部分集合に基づいて生成する。

【０１２２】

クラス判定手段８３（例えば、自己評価部１３）は、分類器が生成される毎に、分類器を用いて、部分集合に属する個々のデータのクラスを判定する。

【０１２３】

類似度算出手段８４（例えば、類似度算出部１４）は、データの集合の部分集合が生成され、クラス判定手段がその部分集合に属する個々のデータのクラスを判定したときに、同一のクラスと判定されたデータ同士の類似度に値を加算する。

【０１２４】

本発明によれば、データの重複を許して、データの集合から部分集合を複数回生成し、部分集合毎に、分類器を生成して、部分集合に属するデータのクラスを判定する。そして、この判定結果を用いて類似度を計算するので、データ中に互いに関連のある属性が多く存在してもそれらの関連のある属性の影響を強く受けることなく、データの類似度を求めることができる。

【０１２５】

また、上記の実施形態には、データ同士の類似度に基づいて、データの集合に属するデータを複数のグループに分類するデータグループ化手段（例えば、類似度クラスタリング部２１）を備える構成が開示されている。そのような構成によれば、データの集合に含まれる個別データを分類することができる。

【０１２６】

また、上記の実施形態には、データグループ化手段が、データの集合に属する個々のデータをそれぞれ別々のグループに分類し、互いに異なる二つのグループに属するデータ同士の類似度を求め、類似度が最大となる二つのグループを併合することを繰り返し、グループの総数を目標数まで減少させる構成が開示されている。そのような構成によれば、データの集合に含まれる個別データを目標数のグループに分類することができる。

【０１２７】

また、上記の実施形態には、データ集合に属するデータを、特定の属性の属性値に応じて、グループに分類する属性データ分類手段（属性データ分類部３１）と、データグループ化手段によって分類されたデータのグループと、属性データ分類手段によって分類されたデータのグループとの関係に基づいて、類似度算出に対する特定の属性の関連度を計算する関連度計算手段（関連性算出部３２）とを備える構成が開示されている。そのような構成によれば、類似度を用いて分類を行った結果と属性値との関連の度合いを調べることができる。

【０１２８】

また、上記の実施形態には、部分集合生成手段８１が、データの集合からデータをランダムサンプリングすることによって、集合の部分集合を生成する構成が開示されている。

【０１２９】

また、上記の実施形態には、類似度算出手段８４が、クラス判定手段８３によって特定のクラスと判定されたデータ同士の類似度に対してのみ値を加算する構成が開示されている。そのような構成によれば、特定のクラスのデータの持つ特徴の違いを重視して類似度を算出できる。

【０１３０】

また、上記の実施形態には、分類器生成手段８２が、部分集合に属するデータのうち所与のクラスが特定のクラスであるデータを加重して分類器を生成する構成が開示されている。そのような構成によれば、得られた類似度に基づいてデータを分類するときに、特定のクラス以外のクラスのデータが持つ特徴の違いを重視してグループに分類することができる。

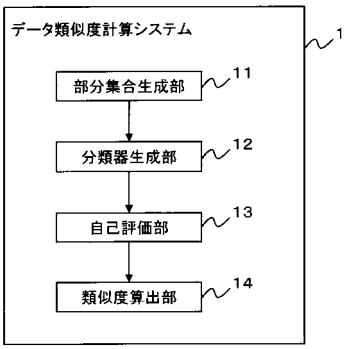
【０１３１】

また、上記の実施形態には、部分集合生成手段８１が、少なくとも顧客の特徴または販売条件を属性とし顧客の行動をクラスとするデータの集合から、その集合の部分集合を生成する構成が開示されている。

【産業上の利用可能性】

【０１３２】
本発明は、データの集合に属する各データ間の類似度を求めるデータ類似度計算システムや、データ間の類似度を計算してデータを分類する分類システムに好適に適用される。
【図面の簡単な説明】
【０１３３】
【図１】本発明の第１の実施形態の例を示すブロック図である。
【図２】データの例を示す説明図である。
【図３】第１の実施形態のデータ類似度計算システムの処理経過の例を示すフローチャートである。
【図４】分類器の例を示す説明図である。
【図５】分類器から判定されたクラスの例を示す説明図である。
【図６】本発明の第２の実施形態の例を示すブロック図である。
【図７】最短距離法によって個別データをクラスタリングする処理経過の例を示すフローチャートである。
【図８】本発明の第３の実施形態の例を示すブロック図である。
【図９】属性の関連度を求める動作の例を示すフローチャートである。
【図１０】χ^2 分布表の例を示す説明図である。
【図１１】本発明の概要を示すブロック図である。
【図１２】データの例を示す説明図である。
【符号の説明】
【０１３４】
１１　部分集合生成部
１２　分類器生成部
１３　自己評価部
１４　類似度算出部
２１　類似度クラスタリング部
３１　属性データ分類部
３２　関連性算出部

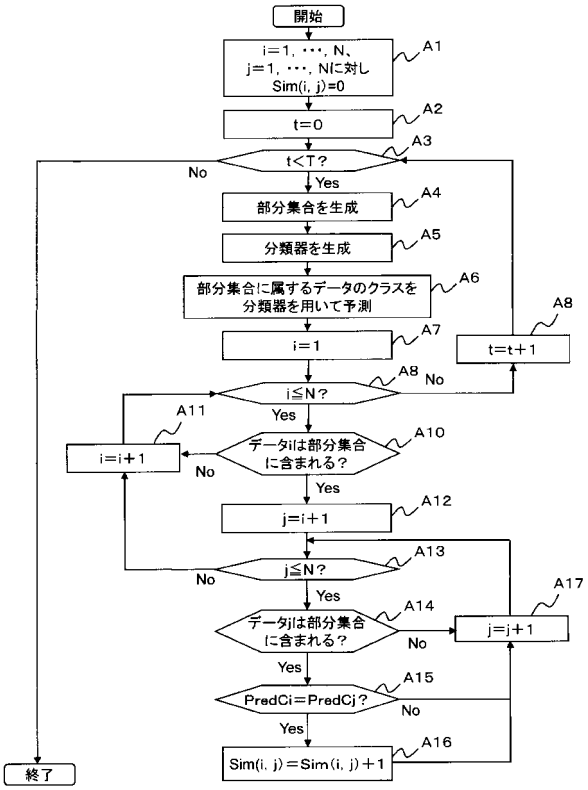
【図１】



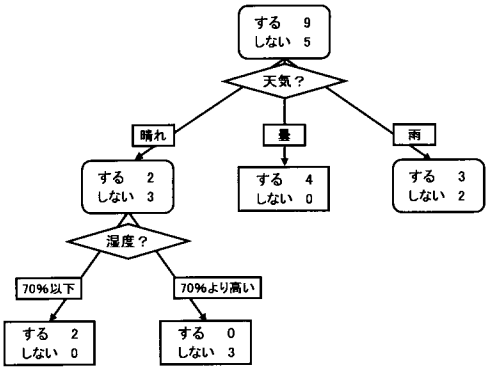
【図２】

属性				クラス
天気	気温(度)	湿度(%)	風が強いか否か	購入するか
晴れ	29	85	強くない	しない
晴れ	27	90	強い	しない
曇	28	78	強くない	する
雨	21	96	強くない	する
雨	20	80	強くない	する
雨	18	70	強い	しない
曇	18	65	強い	する
晴	22	95	強くない	しない
晴	21	70	強くない	する
雨	24	80	強くない	する
晴	24	70	強い	する
曇	22	90	強い	する
曇	27	75	強くない	する
雨	22	80	強い	しない

【図３】



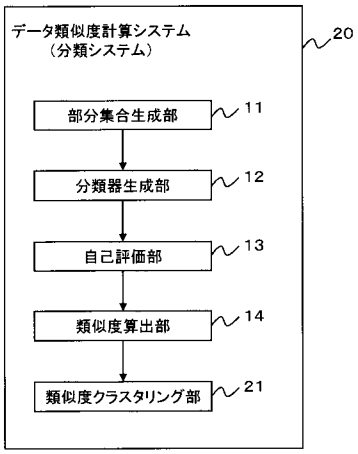
【図４】



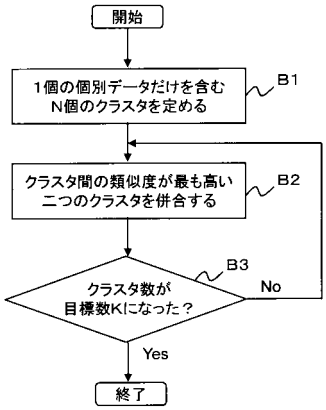
【図５】

属性				クラス
天気	気温(度)	湿度(%)	風が強いか	購入するか
晴れ	29	85	強くない	しない
晴れ	27	90	強い	しない
曇	28	78	強くない	する
雨	21	96	強くない	する
雨	20	80	強くない	する
雨	18	70	強い	する
曇	18	65	強い	する
晴	22	95	強くない	しない
晴	21	70	強くない	する
雨	24	80	強くない	する
晴	24	70	強い	する
曇	22	90	強い	する
曇	27	75	強くない	する
雨	22	80	強い	する

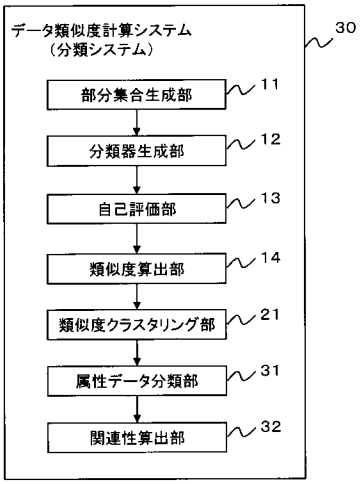
【図６】


データ類似度計算システム
（分類システム）
20
部分集合生成部
11
分類器生成部
12
自己評価部
13
類似度算出部
14
類似度クラスタリング部
21

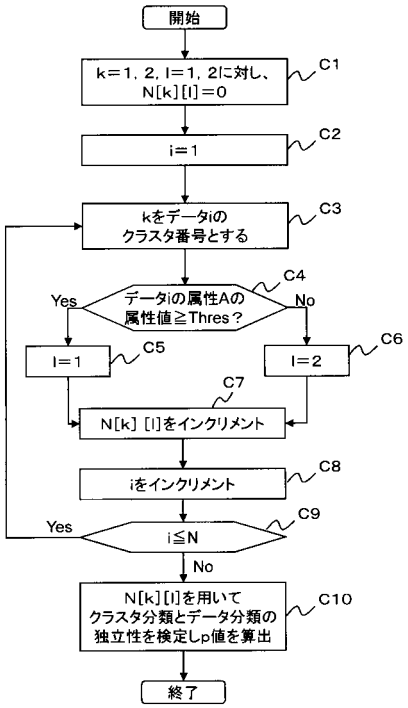
【図７】


開始
1個の個別データだけを含む
N個のクラスタを定める
B1
クラスタ間の類似度が最も高い
二つのクラスタを併合する
B2
クラスタ数が
目標数Kになった？
B3
No
Yes
終了

【図８】


データ類似度計算システム
（分類システム）
30
部分集合生成部
11
分類器生成部
12
自己評価部
13
類似度算出部
14
類似度クラスタリング部
21
属性データ分類部
31
関連性算出部
32

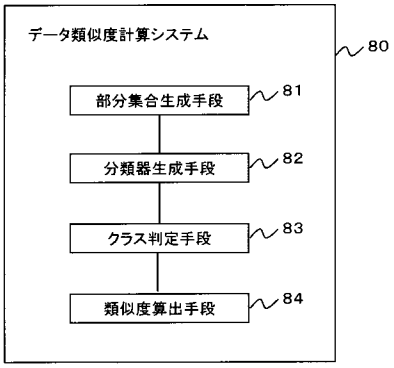
【図９】


開始
k＝1，2，l＝1，2に対し、
N[k][l]＝0
C1
i＝1
C2
kをデータiの
クラスタ番号とする
C3
データiの属性Aの
属性値≧Thres？
C4
Yes
No
l＝1
C5
l＝2
C6
C7
N[k][l]をインクリメント
iをインクリメント
C8
Yes
i≦N
C9
No
N[k][l]を用いて
クラスタ分類とデータ分類の
独立性を検定しp値を算出
C10
終了

【図１０】

	α								
ν	0.7	0.5	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0.15	0.45	1.07	1.64	2.71	3.84	5.41	6.63	10.83
2	0.71	1.39	2.41	3.22	4.61	5.99	7.82	9.21	13.82
3	1.42	2.37	3.66	4.64	6.25	7.81	9.84	11.34	16.27
：									

【図１１】



【図１２】

データ番号	属性ａ	属性ｂ	属性ｃ	属性ｄ	属性ｅ	クラス
1	1	1	1	2	2	A
2	1	1	1	2	2	A
3	1	1	1	2	2	A
4	1	1	1	2	2	A
5	1	2	2	1	1	A
6	2	1	1	2	2	A
7	2	2	2	1	1	A
8	2	2	2	1	1	B
9	2	2	2	1	1	B
１０	2	2	2	1	1	B

フロントページの続き

(56)参考文献 特開２００５－２７５５５６（ＪＰ，Ａ）
特開２００２－１４９６９７（ＪＰ，Ａ）
特開２００５－０７８２４０（ＪＰ，Ａ）
Leo Breiman, RF/tools A Class of Two-eyed Algorithms, Proceedings of SIAM International Conference on Data Mining, 米国, Society for Industrial and Applied Mathematics, ２００３年 5月 2日, Vol.2003(Keynote4), pp.1-56., [ONLINE]取得日2013.03.07, ＵＲＬ, http://oz.berkeley.edu/users/breiman/siamtalk2003.pdf
金 明哲、村上 征勝, ランダムフォレスト法による文章の書き手の同定, 統計数理, 日本, 統計数理研究所, ２００７年, 第５５巻、第２号, pp.255-268., [ONLINE]取得日2013/03/07, ＵＲＬ, http://www.ism.ac.jp/editsec/toukei/pdf/55-2-255.pdf
Andy Liaw,Matthew Wiener, Classification and Regression by randomForest, R News, オーストリア, the R Foundation for Statistical Computing, communications, ２００２年１２月, Vol.2/3, pp.18-22., [ONLINE]取得日2013.03.07, ＵＲＬ, http://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １７／３０
ＪＳＴＰｌｕｓ（ＪＤｒｅａｍＩＩＩ）