(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4255779号
(P4255779)

(45)発行日　平成21年4月15日(2009.4.15)　(24)登録日　平成21年2月6日(2009.2.6)

(51)Int.Cl.		F I		
G06F 19/00	(2006.01)	G06F	19/00	１３０
G06F 17/15	(2006.01)	G06F	17/15	
G06F 17/18	(2006.01)	G06F	17/18	Ｚ
G06Q 50/00	(2006.01)	G06F	17/60	１０６

請求項の数 10　(全 31 頁)

(21)出願番号　特願2003-272648(P2003-272648)
(22)出願日　平成15年7月10日(2003.7.10)
(65)公開番号　特開2005-32117(P2005-32117A)
(43)公開日　平成17年2月3日(2005.2.3)
審査請求日　平成17年8月10日(2005.8.10)

(73)特許権者　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町２２番２２号
(74)代理人　110000338
特許業務法人原謙三国際特許事務所
(74)代理人　100088281
弁理士　田畑　昌男
(72)発明者　竹内　博明
大阪府大阪市阿倍野区長池町２２番２２号
シャープ株式会社内

審査官　小原　正信

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１９／００

(54)【発明の名称】データ分析装置およびデータ分析方法並びにデータ分析プログラム

(57)【特許請求の範囲】

【請求項１】

分析対象データ格納部に格納された、複数の入力属性x_j（$1 \leqq j \leqq N$、Ｎは入力属性の個数）と、１つの出力属性ｙとで構成されるデータの集合である基本データ群ＤＡを分析対象とし、入力属性と出力属性との因果関係を分析するデータ分析装置であって、

基本データ群ＤＡに含まれる文字属性のデータを、一義的な変換ルールに従って数値属性のデータに変換することによって、数値属性のデータの集合である数値型基本データ群ＤＡ０を生成する文字－数値データ変換手段と、

数値型基本データ群ＤＡ０を、数値型基本データ群ＤＡ０に含まれる出力属性ｙの数値と、出力属性ｙの所定閾値との大小関係の比較に基づいて、第１データ群ＤＡ１と、第２データ群ＤＡ２とに分類する分類手段と、

上記複数の入力属性のうちの１つの入力属性x_jについて、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第１データ群ＤＡ１に属するデータの個数の、第１データ群ＤＡ１に属する全てのデータの個数に対する比率である第１の頻度（１－x_j頻度累積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第２データ群ＤＡ２に属するデータの個数の、第２データ群ＤＡ２に属する全てのデータの個数に対する比率である第２の頻度（２－x_j頻度累積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、第１の頻度と第２の頻度との差分（x_j頻度累積差％）を求める演算を、上記複数の入力属性の各々について行なう第１の評価手段と、

上記複数の入力属性のうちの１つの入力属性x_jについて、第１の評価手段で該１つの入力属性x_jのとり得る数値毎に演算された差分（x_j頻度累積差％）に基づいて、最大の差分が求められた数値を当該入力属性x_jの閾値x_{j_th}として決定することを、上記複数の入力属性の各々について行なう閾値決定手段と、

上記複数の入力属性のうちの１つの入力属性x_jについて、閾値決定手段にて決定された該入力属性x_jの閾値x_{j_th}における、第１の頻度（１－x_j頻度累積％）に対する第２の頻度（２－x_j頻度累積％）の比率である第１の比率と、閾値決定手段にて決定された該入力属性x_jの閾値x_{j_th}における、（１００％－第１の頻度（１－x_j頻度累積％））に対する（１００％－第２の頻度（２－x_j頻度累積％））の比率である第２の比率とを演算するとともに、第１の比率および第２の比率のうちの大きい方の比率を選択することを、上記複数の入力属性の各々について行なう第２の評価手段と、

上記第２の評価手段にて入力属性毎に選択された比率のうち、最も大きい比率を持つ入力属性x_j、該入力属性x_jの閾値x_{j_th}、および該最も大きい比率が第１の比率および第２の比率の何れであるかを示す種別を、入力属性条件を示すデータとして抽出するとともに、当該入力属性条件を分析結果データ格納部に格納する要因抽出手段とを含むことを特徴とするデータ分析装置。

【請求項２】

上記要因抽出手段で抽出された入力属性条件に基づいて、数値型基本データ群ＤＡ０を、上記入力属性条件を満たす要因データ群と上記入力属性条件を満たさない他データ群とに分割し、分類されたデータ群のうちの少なくとも一方を新たな数値型基本データ群ＤＡ０として分類手段に送る分割手段をさらに含み、

分類手段による処理、第１の評価手段による処理、閾値決定手段による処理、第２の評価手段による処理、要因抽出手段による処理、および分割手段による処理からなる一連の処理が繰り返し実行されるようになっていることを特徴とする請求項１に記載のデータ分析装置。

【請求項３】

上記分割手段は、分類されたデータ群のうち他データ群のみを選択して新たな数値型基本データ群ＤＡ０として分類手段に送るものであることを特徴とする請求項２に記載のデータ分析装置。

【請求項４】

終了条件を満たしているかを判定する終了条件判定手段をさらに含み、上記終了条件判定手段において終了条件を満たしていると判定されると、上記一連の処理の実行を終了するようになっていることを特徴とする請求項２に記載のデータ分析装置。

【請求項５】

上記終了条件判定手段は、分類手段で分類された第２データ群のデータ数が０であるかを終了条件として判定を行なうことを特徴とする請求項４に記載のデータ分析装置。

【請求項６】

予め定められた設定情報に従って、または、使用者からの入力に応じて、出力属性の上記所定閾値を設定する閾値設定手段をさらに含むことを特徴とする請求項１または２に記載のデータ分析装置。

【請求項７】

上記入力属性は、製品の製造工程における製造プロセス条件および／またはインライン検査結果であり、

上記出力属性は、製品の品質判定結果であり、

上記第２データ群は、品質判定結果が不良のデータ群であることを特徴とする請求項１または２に記載のデータ分析装置。

【請求項８】

請求項１に記載のデータ分析装置を用いて、分析対象データ格納部に格納された、複数の入力属性x_j（$1 \leqq j \leqq N$、Ｎは入力属性の個数）と、１つの出力属性ｙとで構成されるデータの集合である基本データ群ＤＡを分析対象とし、入力属性と出力属性との因果関

係を分析するデータ分析方法であって、
上記文字－数値データ変換手段により、基本データ群ＤＡに含まれる文字属性のデータを、一義的な変換ルールに従って数値属性のデータに変換することによって、数値属性のデータの集合である数値型基本データ群ＤＡ０を生成する文字－数値データ変換ステップと、
上記分類手段により、数値型基本データ群ＤＡ０を、数値型基本データ群ＤＡ０に含まれる出力属性ｙの数値と、出力属性ｙの所定閾値との大小関係の比較に基づいて、第１データ群ＤＡ１と、第２データ群ＤＡ２とに分類する分類ステップと、
上記第１の評価手段により、上記複数の入力属性のうちの１つの入力属性$ｘ_{j}$について、該１つの入力属性$ｘ_{j}$のとり得る数値毎に、当該数値以下の数値を持つデータのうち、第１データ群ＤＡ１に属するデータの個数の、第１データ群ＤＡ１に属する全てのデータの個数に対する比率である第１の頻度（１－$ｘ_{j}$頻度累積％）を求める演算を行い、かつ、該１つの入力属性$ｘ_{j}$のとり得る数値毎に、当該数値以下の数値を持つデータのうち、第２データ群ＤＡ２に属するデータの個数の、第２データ群ＤＡ２に属する全てのデータの個数に対する比率である第２の頻度（２－$ｘ_{j}$頻度累積％）を求める演算を行い、かつ、該１つの入力属性$ｘ_{j}$のとり得る数値毎に、第１の頻度と第２の頻度との差分（$ｘ_{j}$頻度累積差％）を求める演算を、上記複数の入力属性の各々について行なう第１の評価ステップと、
上記閾値決定手段により、上記複数の入力属性のうちの１つの入力属性$ｘ_{j}$について、第１の評価手段で該１つの入力属性$ｘ_{j}$のとり得る数値毎に演算された差分（$ｘ_{j}$頻度累積差％）に基づいて、最大の差分が求められた数値を当該入力属性$ｘ_{j}$の閾値$ｘ_{j-th}$として決定することを、上記複数の入力属性の各々について行なう閾値決定ステップと、
上記第２の評価手段により、上記複数の入力属性のうちの１つの入力属性$ｘ_{j}$について、閾値決定手段にて決定された該入力属性$ｘ_{j}$の閾値$ｘ_{j-th}$における、第１の頻度（１－$ｘ_{j}$頻度累積％）に対する第２の頻度（２－$ｘ_{j}$頻度累積％）の比率である第１の比率と、閾値決定手段にて決定された該入力属性$ｘ_{j}$の閾値$ｘ_{j-th}$における、（１００％－第１の頻度（１－$ｘ_{j}$頻度累積％））に対する（１００％－第２の頻度（２－$ｘ_{j}$頻度累積％））の比率である第２の比率とを演算するとともに、第１の比率および第２の比率のうちの大きい方の比率を選択することを、上記複数の入力属性の各々について行なう第２の評価ステップと、
上記要因抽出手段により、上記第２の評価手段にて入力属性毎に選択された比率のうち、最も大きい比率を持つ入力属性$ｘ_{j}$、該入力属性$ｘ_{j}$の閾値$ｘ_{j-th}$、および該最も大きい比率が第１の比率および第２の比率の何れであるかを示す種別を、（補正前請求項７、段落００８２）入力属性条件を示すデータとして抽出するとともに、当該入力属性条件を分析結果データ格納部に格納する要因抽出ステップとを含むことを特徴とするデータ分析方法。

【請求項９】
分析対象データ格納部に格納された、複数の入力属性$ｘ_{j}$（１≦ｊ≦Ｎ、Ｎは入力属性の個数）と、１つの出力属性ｙとで構成されるデータの集合である基本データ群ＤＡを分析対象とし、入力属性と出力属性との因果関係を分析するデータ分析装置が備えるコンピュータを機能させるためのデータ分析プログラムであって、
上記データ分析装置は、
基本データ群ＤＡに含まれる文字属性のデータを、一義的な変換ルールに従って数値属性のデータに変換することによって、数値属性のデータの集合である数値型基本データ群ＤＡ０を生成する文字－数値データ変換手段と、
数値型基本データ群ＤＡ０を、数値型基本データ群ＤＡ０に含まれる出力属性ｙの数値と、出力属性ｙの所定閾値との大小関係の比較に基づいて、第１データ群ＤＡ１と、第２データ群ＤＡ２とに分類する分類手段と、
上記複数の入力属性のうちの１つの入力属性$ｘ_{j}$について、該１つの入力属性$ｘ_{j}$のとり得る数値毎に、当該数値以下の数値を持つデータのうち、第１データ群ＤＡ１に属する

データの個数の、第１データ群ＤＡ１に属する全てのデータの個数に対する比率である第１の頻度（１－ｘ$_j$頻度累積％）を求める演算を行い、かつ、該１つの入力属性ｘ$_j$のとり得る数値毎に、当該数値以下の数値を持つデータのうち、第２データ群ＤＡ２に属するデータの個数の、第２データ群ＤＡ２に属する全てのデータの個数に対する比率である第２の頻度（２－ｘ$_j$頻度累積％）を求める演算を行い、かつ、該１つの入力属性ｘ$_j$のとり得る数値毎に、第１の頻度と第２の頻度との差分（ｘ$_j$頻度累積差％）を求める演算を、上記複数の入力属性の各々について行なう第１の評価手段と、

上記複数の入力属性のうちの１つの入力属性ｘ$_j$について、第１の評価手段で該１つの入力属性ｘ$_j$のとり得る数値毎に演算された差分（ｘ$_j$頻度累積差％）に基づいて、最大の差分が求められた数値を当該入力属性ｘ$_j$の閾値ｘ$_{j-th}$として決定することを、上記複数の入力属性の各々について行なう閾値決定手段と、

上記複数の入力属性のうちの１つの入力属性ｘ$_j$について、閾値決定手段にて決定された該入力属性ｘ$_j$の閾値ｘ$_{j-th}$における、第１の頻度（１－ｘ$_j$頻度累積％）に対する第２の頻度（２－ｘ$_j$頻度累積％）の比率である第１の比率と、閾値決定手段にて決定された該入力属性ｘ$_j$の閾値ｘ$_{j-th}$における、（１００％－第１の頻度（１－ｘ$_j$頻度累積％））に対する（１００％－第２の頻度（２－ｘ$_j$頻度累積％））の比率である第２の比率とを演算するとともに、第１の比率および第２の比率のうちの大きい方の比率を選択することを、上記複数の入力属性の各々について行なう第２の評価手段と、

上記第２の評価手段にて入力属性毎に選択された比率のうち、最も大きい比率を持つ入力属性ｘ$_j$、該入力属性ｘ$_j$の閾値ｘ$_{j-th}$、および該最も大きい比率が第１の比率および第２の比率の何れであるかを示す種別を、入力属性条件を示すデータとして抽出するとともに、当該入力属性条件を分析結果データ格納部に格納する要因抽出手段とを含み、

コンピュータを上記の各手段として機能させるためのデータ分析プログラム。

【請求項１０】

請求項９に記載のデータ分析プログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、分析対象である出力属性（目的属性）、例えば製造工程で製造される製品の特性等と、出力属性に影響を与える属性である入力属性（説明属性）、例えば製造プロセス条件等との因果関係を分析するデータ分析装置およびデータ分析方法並びにデータ分析プログラムに関する。

【背景技術】

【０００２】

出力属性と入力属性との因果関係を分析する有効な手法としては、決定木手法が知られている（特許文献１参照）。この手法では、各入力属性の値で順次切り分けた葉の部分で、出力属性の値がうまくまとまるような木構造を作成する。

【０００３】

図１０は、特許文献１の従来技術の項（特許文献１の段落［０００２］～［０００５］および図２２参照）に記載されている決定木の１例であり、表１のデータ群を分析対象としている。表１のデータ群は、ｘ１，ｘ２，ｘ３，ｘ４の４つの入力属性の値と、これら入力属性に対する出力属性ｙの値とを組とするデータを１２個集めた集合である。この手法で作成される決定木（以下、「従来の決定木－１」と呼ぶ事にする）では、図１０に示すように、出力属性ｙの値Ｘ，Ｙ，Ｚが入力属性ｘ１，ｘ２，ｘ３の各値によって、うまく切り分けられている。

【０００４】

しかし、図１０の従来の決定木－１の作成においては、データを分類する際に、入力属性がとる値の数（属性値の種類数）だけのデータ集合に分類される。例えば、入力属性ｘ２は４種類の値（ａ，ｂ，ｃ，ｄ）をとるので、入力属性ｘ２

による分類により４つの集合に分類される。そのため、入力属性がとる値の数が増えると、決定木が煩雑になる可能性がある。

【０００５】

この課題の解決策として、特許文献１では、各属性において、まとめられる属性値を１つのラベルで表現し、ラベルによりデータ分類する決定木を提案している。

【０００６】

図１１は、特許文献１の実施例（特許文献１の段落［００１０］～［００２８］および図１３参照）に記載のラベル階層である。この実施例では、例えば、４種の属性値（１，２，３，４）からなるｘ３属性について、ｘ３属性値「１」「２」に「２．５以下」というラベルをつけおよび、ｘ３属性値「３」「４」に「２．５以上」というラベルをつけて階層構造を表現している。このラベル階層構造を用いて作成される決定木（以下、この決定木を従来の決定木－２と呼ぶ事にする）は、図１２（特許文献１の段落［００１０］～［００２８］および図１４参照）に示す如くであり、図１０に示す従来の決定木－１に比べて、非常に簡潔である。

【特許文献１】特開平８－３１４７２５号公報（公開日：平成８年(1996)１１月２９日）

【発明の開示】

【発明が解決しようとする課題】

【０００７】

上記従来の決定木生成手法をデバイス等の製品の製造工程における製品特性不良の要因分析に応用する場合を題材にして、従来技術の課題を説明する。

【０００８】

いま、表１の入力属性ｘ１，ｘ２，ｘ３，ｘ４が製品製造工程における各種のプロセスデータやインライン検査データ、出力属性ｙが製造された製品の特性データであり、出力属性ｙ＝Ｙが製品特性不良に相当するものとする。そして、プロセス技術者が、製品特性不良ｙ＝Ｙに対し、特許文献１の従来技術に記載された手法で生成された決定木－１（図１０）、または特許文献１に記載された手法で生成された従来の決定木－２（図１２）を用いて、製品特性不良の要因を調査するものとする。

【０００９】

このとき、特許文献１の従来技術に記載された手法で生成された決定木－１では、注目すべきｙ＝Ｙが樹形の中の複数箇所（図１０の例では４箇所）に分散しているため煩雑であり、「どの入力属性がどの値の範囲にあるから製品特性が悪いのか？」という製品特性不良の要因をプロセス技術者が判断しにくい。図１０の例では、入力属性が４属性だけでかつ各属性値の種類も４つだけであるため、何とか、プロセス技術者が製品特性不良の要因を判断することも可能である。しかしながら、実際のデバイス（特に半導体デバイス）のような製品の製造現場では、１工程につき１０～１００属性程度のプロセスデータやインライン検査データがあり、しかも、その値は多値で非常に広い範囲で分布している。さらに、外乱（入力属性として検出できていない属性）の影響により、各入力属性の値が同じであっても、出力属性の値がばらつく事も多い。これらのような場合に特許文献１の従来技術に記載された手法を用いると、厳密な分析を目指すがあまり、無限数のデータ集合に分類されてしまい、もはや、プロセス技術者が、適正に製品特性不良の要因を特定する事ができなくなる。

【００１０】

一方、特許文献１に開示された手法により生成される決定木－２（図１２）では、ラベル階層による分類がなされているので、決定木が簡潔である。そのため、プロセス技術者が、ｙ＝Ｙなる製品特性不良の要因を特定しやすい。

【００１１】

しかし、この図１２に示す簡潔な決定木－２を作成するには、図１１に示すラベル階層構造を予め定義しておく必要がある。そのため、特許文献１の決定木生成手法は、まとめられる属性値の見当がつかない場合には適用できない。上述したように、実際のデバイスのような製品の製造現場では、１工程につき１０～１００属性程度の、プロセスデータやインライン検査データがあり、しかも、その値は多値で非常に広い範囲で分布している。さらに、外乱（入力属性として検出できていない属性）の影響により、各入力属性の値が同じであっても、出力属性の値がばらつく事も多い。これらのような状況下で、各入力属性に対し、一つのラベルとしてまとめられる属性値を見出す事は、経験豊富なプロセス技術者であっても、非常に困難である。

【００１２】

本発明は、上記従来の問題点を鑑みてなされたものであり、その目的は、ラベル階層構造を予め定義する事なく、簡潔な形で、出力属性と入力属性との因果関係を導き出せるデータ分析装置およびデータ分析方法並びにデータ分析プログラムを提供する事にある。

【課題を解決するための手段】

【００１３】

本発明に係るデータ分析装置は、上記の課題を解決するために、分析対象データ格納部に格納された、複数の入力属性x_j（$1 \leqq j \leqq N$、Ｎは入力属性の個数）と、１つの出力属性ｙとで構成されるデータの集合である基本データ群ＤＡを分析対象とし、入力属性と出力属性との因果関係を分析するデータ分析装置であって、基本データ群ＤＡに含まれる文字属性のデータを、一義的な変換ルールに従って数値属性のデータに変換することによって、数値属性のデータの集合である数値型基本データ群ＤＡ０を生成する文字－数値データ変換手段と、数値型基本データ群ＤＡ０を、数値型基本データ群ＤＡ０に含まれる出力属性ｙの数値と、出力属性ｙの所定閾値との大小関係の比較に基づいて、第１データ群ＤＡ１と、第２データ群ＤＡ２とに分類する分類手段と、上記複数の入力属性のうちの１つの入力属性x_jについて、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第１データ群ＤＡ１に属するデータの個数の、第１データ群ＤＡ１に属する全てのデータの個数に対する比率である第１の頻度（１－x_j頻度累積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第２データ群ＤＡ２に属するデータの個数の、第２データ群ＤＡ２に属する全てのデータの個数に対する比率である第２の頻度（２－x_j頻度累積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、第１の頻度と第２の頻度との差分（x_j頻度累積差％）を求める演算を、上記複数の入力属性の各々について行なう第１の評価手段と、上記複数の入力属性のうちの１つの入力属性x_jについて、第１の評価手段で該１つの入力属性x_jのとり得る数値毎に演算された差分（x_j頻度累積差％）に基づいて、最大の差分が求められた数値を当該入力属性x_jの閾値x_{j-th}として決定することを、上記複数の入力属性の各々について行なう閾値決定手段と、上記複数の入力属性のうちの１つの入力属性x_jについて、閾値決定手段にて決定された該入力属性x_jの閾値x_{j-th}における、第１の頻度（１－x_j頻度累積％）に対する第２の頻度（２－x_j頻度累積％）の比率である第１の比率と、閾値決定手段にて決定された該入力属性x_jの閾値x_{j-th}における、（１００％－第１の頻度（１－x_j頻度累積％））に対する（１００％－第２の頻度（２－x_j頻度累積％））の比率である第２の比率とを演算するとともに、第１の比率および第２の比率のうちの大きい方の比率を選択することを、上記複数の入力属性の各々について行なう第２の評価手段と、上記第２の評価手段にて入力属性毎に選択された比率のうち、最も大きい比率を持つ入力属性x_j、該入力属性x_jの閾値x_{j-th}、および該最も大きい比率が第１の比率および第２の比率の何れであるかを示す種別を、入力属性条件を示すデータとして抽出するとともに、当該入力属性条件を分析結果データ格納部に格納する要因抽出手段とを含むことを特徴としている。

【００１４】

本発明に係るデータ分析方法は、上記の課題を解決するために、前記のデータ分析装置を用いて、分析対象データ格納部に格納された、複数の入力属性x_j（$1 \leqq j \leqq N$、Ｎは入力属性の個数）と、１つの出力属性ｙとで構成されるデータの集合である基本データ群ＤＡを分析対象とし、入力属性と出力属性との因果関係を分析するデータ分析方法であって、上記文字－数値データ変換手段により、基本データ群ＤＡに含まれる文字属性のデータを、一義的な変換ルールに従って数値属性のデータに変換することによって、数値属性のデータの集合である数値型基本データ群ＤＡ０を生成する文字－数値データ変換ステップと、上記分類手段により、数値型基本データ群ＤＡ０を、数値型基本データ群ＤＡ０に含まれる出力属性ｙの数値と、出力属性ｙの所定閾値との大小関係の比較に基づいて、第１データ群ＤＡ１と、第２データ群ＤＡ２とに分類する分類ステップと、上記第１の評価手段により、上記複数の入力属性のうちの１つの入力属性x_jについて、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第１データ群ＤＡ１に属するデータの個数の、第１データ群ＤＡ１に属する全てのデータの個数に対する比率である第１の頻度（１－x_j頻度累積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第２データ群ＤＡ２に属するデータの個数の、第２データ群ＤＡ２に属する全てのデータの個数に対する比率である第２の頻度（２－x_j頻度累積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、第１の頻度と第２の頻度との差分（x_j頻度累積差％）を求める演算を、上記複数の入力属性の各々について行なう第１の評価ステップと、上記閾値決定手段により、上記複数の入力属性のうちの１つの入力属性x_jについて、第１の評価手段で該１つの入力属性x_jのとり得る数値毎に演算された差分（x_j頻度累積差％）に基づいて、最大の差分が求められた数値を当該入力属性x_jの閾値x_{j_th}として決定することを、上記複数の入力属性の各々について行なう閾値決定ステップと、上記第２の評価手段により、上記複数の入力属性のうちの１つの入力属性x_jについて、閾値決定手段にて決定された該入力属性x_jの閾値x_{j_th}における、第１の頻度（１－x_j頻度累積％）に対する第２の頻度（２－x_j頻度累積％）の比率である第１の比率と、閾値決定手段にて決定された該入力属性x_jの閾値x_{j_th}における、（１００％－第１の頻度（１－x_j頻度累積％））に対する（１００％－第２の頻度（２－x_j頻度累積％））の比率である第２の比率とを演算するとともに、第１の比率および第２の比率のうちの大きい方の比率を選択することを、上記複数の入力属性の各々について行なう第２の評価ステップと、上記要因抽出手段により、上記第２の評価手段にて入力属性毎に選択された比率のうち、最も大きい比率を持つ入力属性x_j、該入力属性x_jの閾値x_{j_th}、および該最も大きい比率が第１の比率および第２の比率の何れであるかを示す種別を、入力属性条件を示すデータとして抽出するとともに、当該入力属性条件を分析結果データ格納部に格納する要因抽出ステップとを含むことを特徴としている。

【００１５】

本発明に係るデータ分析プログラムは、上記の課題を解決するために、分析対象データ格納部に格納された、複数の入力属性x_j（$1 \leqq j \leqq N$、Ｎは入力属性の個数）と、１つの出力属性ｙとで構成されるデータの集合である基本データ群ＤＡを分析対象とし、入力属性と出力属性との因果関係を分析するデータ分析装置が備えるコンピュータを機能させるためのデータ分析プログラムであって、上記データ分析装置は、基本データ群ＤＡに含まれる文字属性のデータを、一義的な変換ルールに従って数値属性のデータに変換することによって、数値属性のデータの集合である数値型基本データ群ＤＡ０を生成する文字－数値データ変換手段と、数値型基本データ群ＤＡ０を、数値型基本データ群ＤＡ０に含まれる出力属性ｙの数値と、出力属性ｙの所定閾値との大小関係の比較に基づいて、第１データ群ＤＡ１と、第２データ群ＤＡ２とに分類する分類手段と、上記複数の入力属性のうちの１つの入力属性x_jについて、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第１データ群ＤＡ１に属するデータの個数の、第１データ群ＤＡ１に属する全てのデータの個数に対する比率である第１の頻度（１－x_j頻度累

積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、当該数値以下の数値を持つデータのうち、第２データ群ＤＡ２に属するデータの個数の、第２データ群ＤＡ２に属する全てのデータの個数に対する比率である第２の頻度（２－x_j頻度累積％）を求める演算を行い、かつ、該１つの入力属性x_jのとり得る数値毎に、第１の頻度と第２の頻度との差分（x_j頻度累積差％）を求める演算を、上記複数の入力属性の各々について行なう第１の評価手段と、上記複数の入力属性のうちの１つの入力属性x_jについて、第１の評価手段で該１つの入力属性x_jのとり得る数値毎に演算された差分（x_j頻度累積差％）に基づいて、最大の差分が求められた数値を当該入力属性x_jの閾値x_{j-th}として決定することを、上記複数の入力属性の各々について行なう閾値決定手段と、上記複数の入力属性のうちの１つの入力属性x_jについて、閾値決定手段にて決定された該入力属性x_jの閾値x_{j-th}における、第１の頻度（１－x_j頻度累積％）に対する第２の頻度（２－x_j頻度累積％）の比率である第１の比率と、閾値決定手段にて決定された該入力属性x_jの閾値x_{j-th}における、（１００％－第１の頻度（１－x_j頻度累積％））に対する（１００％－第２の頻度（２－x_j頻度累積％））の比率である第２の比率とを演算するとともに、第１の比率および第２の比率のうちの大きい方の比率を選択することを、上記複数の入力属性の各々について行なう第２の評価手段と、上記第２の評価手段にて入力属性毎に選択された比率のうち、最も大きい比率を持つ入力属性x_j、該入力属性x_jの閾値x_{j-th}、および該最も大きい比率が第１の比率および第２の比率の何れであるかを示す種別を、入力属性条件を示すデータとして抽出するとともに、当該入力属性条件を分析結果データ格納部に格納する要因抽出手段とを含み、コンピュータを上記の各手段として機能させるためのデータ分析プログラムであることを特徴としている。

【００１６】

本発明に係るコンピュータ読み取り可能な記録媒体は、上記の課題を解決するために、上記のデータ分析プログラムを記録したものであることを特徴としている。

【００１７】

上記装置、方法、プログラム、あるいは記録媒体によれば、ラベル階層構造を予め定義する事なく、簡潔な形で、第２データ群に対応する出力属性条件（結果）の要因を抽出できる。それゆえ、例えば第２データ群が悪い結果（例えば不良品の発生）に対応するデータ群であれば、その悪い結果の要因をユーザが容易に把握できる。逆に、第２データ群が良い結果（例えば優れた特性を持つ製品の発生）に対応するデータ群であれば、その良い結果の要因をユーザが容易に把握できる。

【００１８】

本発明に係るデータ分析方法は、上記要因抽出手段で抽出された入力属性条件に基づいて、数値型基本データ群ＤＡ０を、上記入力属性条件を満たす要因データ群と上記入力属性条件を満たさない他データ群とに分割し、分類されたデータ群のうちの少なくとも一方を新たな数値型基本データ群ＤＡ０として分類手段に送る分割手段をさらに含み、分類手段による処理、第１の評価手段による処理、閾値決定手段による処理、第２の評価手段による処理、要因抽出手段による処理、および分割手段による処理からなる一連の処理が繰り返し実行されるようになっていることがより好ましい。

【００１９】

上記構成によれば、複数の要因を節点として木構造を作成できる。それゆえ、単独の相関ルールでは表現し難い複数の要因の絡み合った分析対象であっても、十分高い精度で要因を究明できる。

【００２０】

本発明に係るデータ分析装置は、終了条件を満たしているかを判定する終了条件判定手段をさらに含み、上記終了条件判定手段において終了条件を満たしていると判定されると、上記一連の処理の実行を終了するようになっていることがよ

り好ましい。これにより、必要以上の無駄な処理が行われることを回避できる。
【００２１】
上記第１の評価手段は、各入力属性の全ての数値について、第１データ群中における入力属性がその数値以下であるデータの割合を第１の頻度として演算すると共に、第２データ群中における入力属性がその数値以下であるデータの割合を第２の頻度として演算する頻度演算手段と、各入力属性の全ての数値について、第１の頻度と第２の頻度との差分を演算する差分演算手段とを含むことがより好ましい。これにより、閾値評価指標を容易に演算することができる。
【００２２】
上記第２の評価手段は、第１のルール評価値として、第１データ群中における入力属性が閾値以下であるデータの割合に対する、第２データ群中における入力属性が閾値以下であるデータの割合の比率を第１の比率として演算すると共に、第２のルール評価値として、第１データ群中における入力属性が閾値を超えるデータの割合に対する、第２データ群中における入力属性が閾値を超えるデータの割合の比率を第２の比率として演算し、双方の比率のうち大きい方の比率を抽出するものであり、上記要因抽出手段は、上記第２の評価手段で抽出された、各入力属性の比率のうちで、その値が最大となる、入力属性、該入力属性の閾値、および抽出された比率の種別を上記入力属性条件を示すデータとして抽出するものであることがより好ましい。これにより、第１および第２のルール評価値を容易に演算することができる。
【発明の効果】
【００２３】
本発明の装置、方法、プログラム、記録媒体によれば、以上のように、ラベル階層構造を予め定義する事なく、「入力属性が閾値以下」あるいは「入力属性が閾値を超える」といった非常に簡潔な形で、問題事象である特定の出力属性条件（問題事象）が発生する要因を導き出すことが可能となる。また、複数の要因を導き出せば、それぞれの要因（入力属性）における「入力属性が閾値以下」あるいは「入力属性が閾値を超える」といった条件の組み合わせによる非常に簡潔な形の決定木として、問題事象に関わる因果関係を導き出せる。
【発明を実施するための最良の形態】
【００２４】
本発明の一実施形態を以下に説明する。
【００２５】
まず、本実施形態のデータ分析装置を図１に基づいて説明する。
【００２６】
図１に示すように、データ分析装置は、文字－数値データ変換部１、分析対象データ格納部２、閾値設定部（閾値設定手段）３、データ分類部（分類手段）４、データ列抽出部５、頻度演算部（第１の評価手段、頻度演算手段）６、頻度累積差演算部（第１の評価手段、差分演算手段）７、入力属性閾値決定部（閾値決定手段）８、頻度累積比率演算部（第２の評価手段）１６、要因抽出部（要因抽出手段）９、要因未発見データ抽出部（分割手段）１０、終了条件判定部（終了条件判定手段）１１、入力属性閾値テーブル作成部１２、寄与率演算部１３、分析結果データ格納部１４、および出力部１５を備えている。
【００２７】
次に、次の表１のデータ群ＤＡを分析対象とする場合を例にとって、本実施形態のデータ分析方法を図２に基づいて説明する。表１のデータ群ＤＡは、ハードディスク等の格納部２に格納されている。
【００２８】

【表１】

データ群DA

id	x1	x2	x3	x4	y
1	A	a	1	20	X
2	C	a	2	20	X
3	B	b	1	30	X
4	D	b	1	40	X
5	C	a	3	10	Y
6	D	b	3	30	Y
7	A	c	1	10	Y
8	D	d	4	20	Y
9	A	a	4	40	Z
10	B	a	3	30	Z
11	A	b	3	10	Z
12	A	b	4	20	Z

【００２９】

表１のデータ群ＤＡは、１～１２のｉｄ（識別子）を持つ１２個のデータから構成されている。表１において、ｘ１，ｘ２，ｘ３，ｘ４は入力属性である。入力属性ｘ１は４つの文字Ａ，Ｂ，Ｃ，Ｄのいずれかをとる文字属性である。入力属性ｘ２は４つの文字ａ，ｂ，ｃ，ｄのいずれかをとる文字属性である。入力属性ｘ３は４つの離散値１，２，３，４のいずれかをとる離散属性である。入力属性ｘ４は４つの離散値１０，２０，３０，４０のいずれかをとる離散属性である。なお、入力属性は、連続した数値をとる連続属性でもよい。

【００３０】

また、表１において、ｙは出力属性である。出力属性は、文字属性であってもよく、離散属性でもよく、また連続属性でもよいが、ここでは、３つの文字Ｘ，Ｙ，Ｚのいずれかをとる文字属性である。

【００３１】

本実施形態のデータ分析方法では、ｙ＝Ｙなる場合を問題事象として、出力属性ｙがＹとなる要因を分析する。

【００３２】

なお、分析対象データの例としては、例えば、入力属性が、製品の製造工程における製造プロセス条件および／またはインライン検査結果（製造ライン途中での検査結果）、出力属性が製品の品質判定結果、ｙ＝Ｙなる問題事象が品質判定結果の不良であるデータが挙げられる。この場合、本実施形態のデータ分析方法により入力属性と出力属性との因果関係を分析し、ｙ＝Ｙなる問題事象の要因を導き出すことで、デバイス特性不良等の不良品の発生を解消する対策を容易に図ることが可能となる。したがって、歩留まりの向上等のような製造プロセスの改善を容易に図ることが可能となる。

【００３３】

分析対象データのより具体的な例としては、例えば、入力属性ｘ１、ｘ２、ｘ３、ｘ４が、プラズマＣＶＤプロセスの、ガス流量、ガス圧力、投入電力、成膜時間などのプロセスデータで、出力属性ｙが、プラズマＣＶＤプロセスで形成される薄膜の膜厚であるようなデータが挙げられる。また、これら入力属性および出力属性の値は、連続属性でも離散属性でも文字属性でもよい。文字属性の場合には、例えば、出力属性が膜厚の例で、‘大’、‘中’、‘小’といった具合に表現される。

［ステップ０］

まず、文字－数値データ変換部１が、ハードディスク等の分析対象データ格納部２に格納された表１のデータ群ＤＡにおける文字属性を下記の変換ルールに従って数値属性（数値データ）に変換する（Ｓ０）。これにより、各データは、数値データに変換される。そして、文字－数値データ変換部１は、変換されたデー

タ群をデータ分類部４に送る。
（ｘ１）Ａ→１、Ｂ→２、Ｃ→３、Ｄ→４
（ｘ２）ａ→１、ｂ→２、ｃ→３、ｄ→４
（ｘ３）変換せず
（ｘ４）変換せず
（ｙ）Ｘ→１、Ｙ→２、Ｚ→３
この変換ルールは、可能な限り、変換後の入力属性の数値が大きいほど出力属性の数値が大きくなるようにあるいはその逆順となるように設定されることが好ましい。なお、変換ルールは、一義性さえあればよく、上記の例に限られない。

【００３４】

上記変換ルールにて数値データに変換されたデータ群ＤＡ０は、表２に示す通りである。

【００３５】

【表２】

基本データ群DA0

id	x1	x2	x3	x4	y
1	1	1	1	20	1
2	3	1	2	20	1
3	2	2	1	30	1
4	4	2	1	40	1
5	3	1	3	10	2
6	4	2	3	30	2
7	1	3	1	10	2
8	4	4	4	20	2
9	1	1	4	40	3
10	2	1	3	30	3
11	1	2	3	10	3
12	1	2	4	20	3

【００３６】

この変換により、得られたデータ群ＤＡ０は、離散値をとる複数の入力属性（説明属性）と出力属性（目的属性）とで構成されるデータの集合となる。以下、データ群ＤＡ０を基本データ群と呼ぶ事にする。

［ステップ１］

閾値設定部３は、予め定められた設定情報に従って、あるいは使用者が図示しないキーボードやマウス等の入力部から問題事象の属性値ｙ＝Ｙを入力したことに応答して、データ群ＤＡのｙ＝Ｙなる問題事象に対応する基本データ群ＤＡ０の出力属性ｙの閾値（出力属性閾値）y_{th}を設定し、データ分類部４に出力する（Ｓ１）。この例においては、データ群ＤＡのｙ＝Ｙなる問題事象に対応する基本データ群ＤＡ０の出力属性ｙの閾値は、$y_{th}=2$である。

［ステップ２］

次に、データ分類部４が、基本データ群ＤＡ０の出力属性ｙの値と、閾値設定部３から出力された出力属性閾値y_{th}との比較論理（１）（２）に基づいて、基本データ群ＤＡ０を、第１データ群ＤＡ１と第２データ群ＤＡ２とに２分化（分類）する（Ｓ２）。

【００３７】

（１）$y>y_{th}$または$y<y_{th}$→ＤＡ１
（２）$y=y_{th}$→ＤＡ２

言い換えると、データ分類部４は、基本データ群ＤＡ０を、出力属性が出力属性閾値y_{th}と一致しない（すなわち１または３である）第１データ群ＤＡ１と、出力属性が出力属性閾値y_{th}（＝２）と一致する第２データ群ＤＡ２とに分類する。第２データ群ＤＡ２は問題事象（例えば、デバイス特性不良など）のデータ群である。すなわち、第２データ群ＤＡ２は出力属性ｙが問題事象を表す属性値（

２）であるデータ群であり、第１データ群ＤＡ１は出力属性ｙが問題事象を表していない属性値（１または３）であるデータ群である。

【００３８】

第１データ群ＤＡ１を表３に、第２データ群ＤＡ２を表４に示す。

【００３９】

【表３】

第1データ群DA1

id	x1	x2	x3	x4	y
1	1	1	1	20	1
2	3	1	2	20	1
3	2	2	1	30	1
4	4	2	1	40	1
9	1	1	4	40	3
10	2	1	3	30	3
11	1	2	3	10	3
12	1	2	4	20	3

【００４０】

【表４】

第2データ群DA2

id	x1	x2	x3	x4	y
5	3	1	3	10	2
6	4	2	3	30	2
7	1	3	1	10	2
8	4	4	4	20	2

【００４１】

なお、以下では、適宜、第１データ群ＤＡ１を良品（ＯＫ品）データ群、第２データ群ＤＡ２を不良品（ＮＧ品）データ群と呼ぶ事にする。

［ステップ３］

次に、データ列抽出部５が、良品データ群ＤＡ１（表３）から、入力属性ｘｊ（１≦ｊ≦４）の各々のデータ列を抽出する（Ｓ３）。このデータ列を１－ｘｊデータ群と呼ぶ事にする。

【００４２】

同様に、データ列抽出部５は、不良品データ群ＤＡ２（表４）からも、入力属性ｘｊ（１≦ｊ≦４）の各々のデータ列を抽出する（Ｓ３）。このデータ列を２－ｘｊデータ群と呼ぶ事にする。

【００４３】

１－ｘｊデータ群を表５～８に、２－ｘｊデータ群を表９～１２に示す。

【００４４】

【表５】

1－x1データ群

id	x1
1	1
2	3
3	2
4	4
9	1
10	2
11	1
12	1

【００４５】

【表６】

１－x2データ群

id	x2
1	1
2	1
3	2
4	2
9	1
10	1
11	2
12	2

【００４６】

【表７】

１－x3データ群

id	x3
1	1
2	2
3	1
4	1
9	4
10	3
11	3
12	4

【００４７】

【表８】

１－x4データ群

id	x4
1	20
2	20
3	30
4	40
9	40
10	30
11	10
12	20

【００４８】

【表９】

２－x1データ群

id	x1
5	3
6	4
7	1
8	4

【００４９】

【表１０】

２－x2データ群

id	x2
5	1
6	2
7	3
8	4

【００５０】

【表１１】

２－x3データ群

id	x3
5	3
6	3
7	1
8	4

【００５１】

【表１２】

２－x4データ群

id	x4
5	10
6	30
7	10
8	20

【００５２】

［ステップ４］

頻度演算部６は、ステップ３で良品データ群ＤＡ１から抽出された１－ｘｊデータ群の各々、およびステップ３で不良品データ群ＤＡ２から抽出された２－ｘｊデータ群の各々を、入力属性ｘｊの値で昇順に並べ替える。そして、入力属性ｘｊの個々の数値について、第１データ群におけるその数値以下のデータ個数の割合を表す１－ｘｊ頻度累積％と、第２データ群におけるその数値以下のデータ個数の割合を表す２－ｘｊ頻度累積％とを計算する（Ｓ４）。

【００５３】

ここでは、表５～８を入力属性ｘｊの値で昇順に並べ替えた表１３～１６を用い、各行（ｉｄ）のデータについて表中でそのデータの位置以上の位置にあるデータ個数の、第１データ群の全データ数（＝８）に対する割合を１－ｘｊ頻度累積％として計算している。同様に、表９～１２を入力属性ｘｊの値で昇順に並べ替えた表１７～２０を用い、各行（ｉｄ）のデータについて表中でそのデータの位置以上の位置にあるデータ個数の、第２データ群の全データ数（＝４）に対する割合を２－ｘｊ頻度累積％として計算している

ここで計算した１－ｘｊ頻度累積％および２－ｘｊ頻度累積％の値を表１３～２０に示す。

【００５４】

【表１３】

１－x1データ群

id	x1	1-x1頻度累積%
1	1	12.5
9	1	25
11	1	37.5
12	1	50
3	2	62.5
10	2	75
2	3	87.5
4	4	100

【００５５】

【表１４】

１－x2データ群

id	x2	1-x2頻度累積%
1	1	12.5
2	1	25
9	1	37.5
10	1	50
3	2	62.5
4	2	75
11	2	87.5
12	2	100

【００５６】

【表１５】

１－x3データ群

id	x3	1-x3頻度累積%
1	1	12.5
3	1	25
4	1	37.5
2	2	50
10	3	62.5
11	3	75
9	4	87.5
12	4	100

【００５７】

【表１６】

１－x4データ群

id	x4	1-x4頻度累積%
11	10	12.5
1	20	25
2	20	37.5
12	20	50
3	30	62.5
10	30	75
4	40	87.5
9	40	100

【００５８】

【表１７】

２－x1データ群

id	x1	2-x1頻度累積%
7	1	25
5	3	50
6	4	75
8	4	100

【００５９】

【表１８】

２－x2データ群

id	x2	2-x2頻度累積%
5	1	25
6	2	50
7	3	75
8	4	100

【００６０】

【表１９】

２－x3データ群

id	x3	2-x3頻度累積%
7	1	25
5	3	50
6	3	75
8	4	100

【００６１】

【表２０】

２－x4データ群

id	x4	2-x4頻度累積%
5	10	25
7	10	50
8	20	75
6	30	100

【００６２】

なお、上述したステップ３・４では、データ列を抽出し、並び替えを行った後に、１－ｘｊ頻度累積％および２－ｘｊ頻度累積％を計算していたが、データ列の抽出や並び替えを行うことなく直接的に１－ｘｊ頻度累積％および２－ｘｊ頻度累積％を計算してもかまわない。

【００６３】

さらに、頻度演算部６は、１－ｘｊ頻度累積％が計算された良品データ群である１－ｘｊデータ群のテーブルと、２－ｘｊ頻度累積％が計算された不良品データ群である２－ｘｊデータ群のテーブルとを結合する。具体的には、入力属性ｘ１について、表１３と表１７とを結合して表２１のｘ１頻度累積テーブルを、入力属性ｘ２について、表１４と表１８とを結合して表２２のｘ２頻度累積テーブルを、入力属性ｘ３について、表１５と表１９とを結合して表２３のｘ３頻度累積テーブルを、入力属性ｘ４について、表１６と表２０とを結合して表２４のｘ４頻度累積テーブルを、それぞれ作成する。

【００６４】

【表２１】

x1頻度累積テーブル

id	x1	1-x1頻度累積%	2-x1頻度累積%
1	1	12.5	
9	1	25	
11	1	37.5	
12	1	50	
3	2	62.5	
10	2	75	
2	3	87.5	
4	4	100	
7	1		25
5	3		50
6	4		75
8	4		100

【００６５】

【表２２】

x2頻度累積テーブル

id	x2	1-x2頻度累積%	2-x2頻度累積%
1	1	12.5	
2	1	25	
9	1	37.5	
10	1	50	
3	2	62.5	
4	2	75	
11	2	87.5	
12	2	100	
5	1		25
6	2		50
7	3		75
8	4		100

【００６６】

【表２３】

x3頻度累積テーブル

id	x3	1-x3頻度累積%	2-x3頻度累積%
1	1	12.5	
3	1	25	
4	1	37.5	
2	2	50	
10	3	62.5	
11	3	75	
9	4	87.5	
12	4	100	
7	1		25
5	3		50
6	3		75
8	4		100

【００６７】
【表２４】

x4頻度累積テーブル

id	x4	1-x4頻度累積%	2-x4頻度累積%
11	10	12.5	
1	20	25	
2	20	37.5	
12	20	50	
3	30	62.5	
10	30	75	
4	40	87.5	
9	40	100	
5	10		25
7	10		50
8	20		75
6	30		100

【００６８】
さらに、頻度演算部６は、表２１～２４の各々の頻度累積テーブルを、入力属性ｘｊの値で昇順に並べ替える。このとき、１－ｘｊ頻度累積％および２－ｘｊ頻度累積％の空欄には、その直前の値を代入する。また、入力属性ｘｊにおいて同じ値が続いている場合には、上記並べ替えられた最終のデータのみを採用する。こうして、頻度演算部６にて、入力属性ｘｊの各値に対して、良品データ群である第１データ群におけるその数値以下のデータ個数の割合を表す１－ｘｊ頻度累積％（Ａ；第１の頻度）と、不良品データ群である第２データ群におけるその数値以下のデータ個数の割合を表す２－ｘｊ頻度累積％（Ｂ；第２の頻度）との双方が算出される（Ｓ４）。
［ステップ５］
次に、頻度累積差演算部７が、入力属性ｘｊの各値に対して、良品の１－ｘｊ頻度累積（Ａ）と、不良品の２－ｘｊ頻度累積（Ｂ）の差分（＝｜Ａ－Ｂ｜）を計算する（Ｓ５）。この差分値を、ｘｊ頻度累積差（＝｜Ａ－Ｂ｜）と呼ぶ。ｘｊ頻度累積差の計算結果を表２５～表２８に示す。
【００６９】
【表２５】

x1頻度累積テーブル

	x1	1-x1頻度累積%(A)	2-x1頻度累積%(B)	x1頻度累積差%(|A-B|)
	1	50	25	25
x1-th→	2	75	25	50
	3	87.5	50	37.5
	4	100	100	0

【００７０】

【表２６】

x2頻度累積テーブル

x2	1−x2頻度累積％(A)	2−x2頻度累積％(B)	x2頻度累積差％(|A−B|)
1	50	25	25
x2−th→ 2	100	50	50
3	100	75	25
4	100	100	0

【００７１】

【表２７】

x3頻度累積テーブル

x3	1−x3頻度累積％(A)	2−x3頻度累積％(B)	x3頻度累積差％(|A−B|)
1	37.5	25	12.5
x3−th→ 2	50	25	25
3	75	75	0
4	100	100	0

【００７２】

【表２８】

x4頻度累積テーブル

x4	1−x4頻度累積％(A)	2−x4頻度累積％(B)	x4頻度累積差％(|A−B|)
1	12.5	50	37.5
x4−th→ 2	50	75	25
3	75	100	25
4	100	100	0

【００７３】

入力属性ｘｊと、良品の１－ｘｊ頻度累積（Ａ）、不良品の２－ｘｊ頻度累積（Ｂ）、ｘｊ頻度累積差｜Ａ－Ｂ｜との関係を図３～図６に示す。

【００７４】

各数値に対するｘｊ頻度累積差｜Ａ－Ｂ｜は、入力属性ｘｊがその数値以下の範囲と、入力属性ｘｊがその数値を超える範囲との２分化によって、良品の第１データ群ＤＡ１と不良品の第２データ群ＤＡ２とがうまく切り分けられているかを表す指標である。言い換えると、ｘｊ頻度累積差｜Ａ－Ｂ｜は、入力属性がその数値以下であるデータが第１データ群および第２データ群のうちの一方に偏っている度合いを表す閾値評価指標である。

【００７５】

なお、ここでは、閾値評価指標としてｘｊ頻度累積差｜Ａ－Ｂ｜を演算してい

るが、各数値に対する閾値評価指標として、データの偏りの度合いを評価する指標、例えば、情報利得（ゲイン）、情報利得比、Ｇｉｎｉインデックス、平均自乗誤差等を用いてもよい。

［ステップ６］

入力属性閾値決定部８が、各入力属性ｘｊについて、ｘｊの個々の値の中で、ｘｊ頻度累積差｜Ａ－Ｂ｜の値が最大となるときの入力属性ｘｊの値を抽出する（Ｓ６）。この値を、入力属性閾値ｘｊ－ｔｈと呼ぶ事にする。

【００７６】

入力属性閾値ｘｊ－ｔｈは、図３～図６を参照して分かるように、ｘｊ≦ｘｊ－ｔｈの範囲と、ｘｊ＞ｘｊ－ｔｈの範囲との２分化によって、良品の第１データ群ＤＡ１と、不良品の第２データ群ＤＡ２との切分けが最も容易となる入力属性ｘｊの値を示している。

【００７７】

なお、ここでは、複数の入力属性について第３ステップ～第６ステップの処理を一括して行っているが、ｊの値を１からＮまで順次増加させて第３ステップ～該第６ステップの処理を繰り返してもよい。

［ステップ７］

次に、頻度累積比率演算部１６が、ｘｊ＝ｘｊ－ｔｈにおいて、良品の１－ｘｊ頻度累積（Ａ）に対する、不良品の２－ｘｊ頻度累積（Ｂ）の比率を計算する。この比率を、２－ｘｊｔｈ下比率（＝Ｂ／Ａ）と呼ぶ事にする。また、１００から良品の１－ｘｊ頻度累積（Ａ）を引いた値（＝１００－Ａ）に対する、１００から不良品の２－ｘｊ頻度累積（Ｂ）を引いた値（＝１００－Ｂ）の比率を計算する。この比率を、２－ｘｊｔｈ上比率（＝（１００－Ｂ）／（１００－Ａ））と呼ぶ事にする。そして、双方の比率のうちの大きい方の値を表す、２－ｘｊｔｈ比率を抽出する。

【００７８】

ここで、２－ｘｊｔｈ下比率は、「ｘｊ≦ｘｊ－ｔｈ」という入力属性条件により、良品の第１データ群と分離して不良品の第２データ群を検出できる割合を表している。また、２－ｘｊｔｈ上比率は、「ｘｊ＞ｘｊ－ｔｈ」という入力属性条件により、良品の第１データ群と分離して不良品の第２データ群を検出できる割合を表している。

【００７９】

言い換えると、２－ｘｊｔｈ下比率は、「入力属性ｘｊが入力属性閾値ｘｊ－ｔｈ以下であれば第２データ群に含まれるデータである」という第１の相関ルールの確からしさを表す評価値（第１のルール評価値）を表している。また、２－ｘｊｔｈ上比率は、「入力属性ｘｊが入力属性閾値ｘｊ－ｔｈを超えていれば第２データ群に含まれるデータである」という第２の相関ルールの確からしさを表す評価値（第２のルール評価値）を表している。

【００８０】

各入力属性ｘｊに対して抽出された入力属性閾値ｘｊ－ｔｈ、ｘｊ＝ｘｊ－ｔｈにおける、良品の１－ｘｊ頻度累積（Ａ）、不良品の２－ｘｊ頻度累積（Ｂ）、ｘｊ頻度累積差｜Ａ－Ｂ｜、２－ｘｊｔｈ下比率Ｂ／Ａ、２－ｘｊｔｈ上比率（１００－Ｂ）／（１００－Ａ）、２－ｘｊｔｈ比率の各値を表２９に示す。

【００８１】

【表２９】

入力属性 xj	入力属性閾値xj-th	1-xj頻度累積% (A)	2-xj頻度累積% (B)	xj頻度累積差% (|A-B|)	2-xjth下比率 (B/A)	2-xjth上比率 ((100-B)/(100-A))	2-xjth比率
1	2	75	25	50	0.333333	3	3
2	2	100	50	50	0.5	∞	∞
3	2	50	25	25	0.5	1.5	1.5
4	1	12.5	50	37.5	4	0.571429	4

【００８２】

［ステップ８］

要因抽出部９が、ｘ１～ｘ４の入力属性のうち、上記ステップ７の２－ｘｊｔｈ比率が最大となる入力属性を抽出する。これにより、２－ｘｊｔｈ比率が最大となる入力属性と、その閾値、採用した比率の種別（上、下）が第２データ群に対応する出力属性条件の要因（入力属性条件）を示すデータとして抽出されることになる。これは、全ての入力属性に関する前記相関ルールのうちで最も高い２－ｘｊｔｈ下比率または２－ｘｊｔｈ上比率を持つ相関ルールの入力属性条件を示すデータを抽出することに相当する。

【００８３】

なお、ここでは、最大のルール評価値を持つ相関ルールの入力属性を抽出するための指標として２－ｘｊｔｈ比率を演算しているが、最大のルール評価値を持つ相関ルールの入力属性を抽出するための指標として、他の評価指標、例えば、支持率（サポート）、確信度（コンフィデンス）、情報利得（ゲイン）、情報利得比、Ｇｉｎｉインデックス、平均自乗誤差等を用いてもよい。

【００８４】

表２９を参照して、入力属性ｘ２＝ｘ２－ｔｈ＝２のとき、２－ｘ２ｔｈ比率＝２－ｘ２ｔｈ上比率＝∞となっている。これは、入力属性条件「ｘ２＞２」にて、良品の第１データ群ＤＡ１と完全に分離して、不良品の第２データ群ＤＡ２を検出できる事を示しており、この事は、図４を参照すると、より理解しやすい。

【００８５】

上記抽出された、入力属性（＝ｘ２）、該入力属性の値を表す入力属性閾値（＝２）、および採用した比率の種別（＝上）のデータを分析結果データ格納部１４に保存する。

【００８６】

以上のようにして、問題事象（不良品の第２データ群ＤＡ２）の一要因として、「ｘ２＞２」という入力属性条件が抽出された。

［ステップ９］

上記ステップ８にて、問題事象（不良品の第２データ群ＤＡ２）の一要因として、「ｘ２＞２」という入力属性条件が抽出されたので、次に、別の要因を調査する。このため、要因未発見データ抽出部１０が、基本データ群ＤＡ０（表２）を入力属性条件「ｘ２＞２」を満たすデータ群（要因データ群）と、基本データ群ＤＡ０（表２）の中で問題事象の要因をまだ発見できていないデータ群（他データ群）、すなわち入力属性条件「ｘ２≦２」を満たす（入力属性条件「ｘ２＞２」を満たさない）データ群とに分割し、問題事象の要因をまだ発見できていないデータ群を抽出する（表３０）。

【００８７】

【表３０】

基本データ群DA0（2回目）

id	x1	x2	x3	x4	y
1	1	1	1	20	1
2	3	1	2	20	1
3	2	2	1	30	1
4	4	2	1	40	1
5	3	1	3	10	2
6	4	2	3	30	2
9	1	1	4	40	3
10	2	1	3	30	3
11	1	2	3	10	3
12	1	2	4	20	3

【００８８】

要因未発見データ抽出部１０は、抽出されたデータ群を次の（新しい）基本データ群ＤＡ０としてデータ分類部４に送る。

［ステップ１０］

そして、ステップ９で抽出されたデータ群を次の基本データ群ＤＡ０として、終了条件判定部１１で終了条件を満たしていると判定されるまで、上記のステップ２～ステップ９の処理が繰り返される。本実施形態の終了条件判定部１１は、繰返し処理中の上記ステップ２において不良品の第２データ群ＤＡ２のデータ個数が０となった場合を終了条件と判定するようになっている。このように不良品の第２データ群ＤＡ２のデータ個数が０となるまで繰り返し処理を実行することにより、詳細な要因分析結果が得られる。

【００８９】

なお、終了条件は、第２データ群ＤＡ２のデータ個数に基づく他の終了条件、例えば、（１）繰返し処理中の上記ステップ２において第２データ群ＤＡ２のデータ個数が所定数以下となった場合、（２）繰返し処理中の上記ステップ２において第１データ群ＤＡ１のデータ個数に対する第２データ群ＤＡ２のデータ個数の割合が所定割合以下となった場合、（３）繰返し処理中の上記ステップ８において抽出された入力属性条件のルール評価値が所定の値を下回った場合等としてもよい。これらのような終了条件を用いた場合、より簡潔で十分な要因分析結果を得ることができる。さらに、簡潔な要因分析結果を得ることを優先する場合には、終了条件を単に繰返し処理を所定回数行った場合としたり、終了条件判定部１１を省いて、可能な限り繰り返し処理を行うようにしてもよい。

【００９０】

今回の例では、２回目の繰り返し処理中のステップ９で抽出した、要因未発見の、ｘ１≦２のデータ群に不良品のデータ（第２データ群ＤＡ２；ｙ＝２）が含まれていなかったため、繰り返し処理は２回目で（２回目の要因抽出を行った時点で）終了した。

［ステップ１１］

入力属性閾値テーブル作成部１２が、ステップ１０の繰り返し処理毎に抽出された入力属性ｘｊと、入力属性閾値ｘｊ－ｔｈと、採用された比率の種別とを格納した入力属性閾値テーブルを作成する（表３１）。

【００９１】

【表３１】

入力属性閾値テーブル

	入力属性 xj	入力属性閾値xj-th	種別	要因条件
1回目	x2	2	上	x2>2
2回目	x1	2	上	x1>2

【００９２】
入力属性閾値テーブル作成部１２では、必要に応じて、入力属性閾値テーブルにおける入力属性閾値ｘｊ－ｔｈの数値を文字データに変換する。文字データへの変換ルールは、ステップ０の変換の逆変換となるルールであり、下記の通りである。
（ｘ１）１→Ａ、２→Ｂ、３→Ｃ、４→Ｄ
（ｘ２）１→ａ、２→ｂ、３→ｃ、４→ｄ
（ｘ３）変換せず
（ｘ４）変換せず
表３１の入力属性閾値テーブルにおける入力属性閾値ｘｊ－ｔｈを文字データに変換した入力属性閾値テーブルを表３２に示す。
【００９３】
【表３２】

入力属性閾値テーブル

	入力属性 xj	入力属性閾値 xj-th	種別	要因条件
1回目	x2	b	上	x2=c or d
2回目	x1	B	上	x1=C or D

【００９４】
この入力属性閾値テーブルは、特許文献１に記載の従来の決定木－２（図１２）において、出力属性ｙ＝Ｙ（ｙ＝２）の切分けに着目した場合の決定木の分類条件に対応する。
［ステップ１２］
次に、寄与率演算部１３が、表３１の入力属性閾値テーブルから、抽出された入力属性の、問題事象（ｙ＝２：不良品データ群である、元の第２データ群ＤＡ２）に対する寄与率（相関ルールの評価指標であるサポートに相当する）を求める。
【００９５】
表３３は、問題事象（不良品）である元の第２データ群ＤＡ２（表４）において、その要因として１回目に抽出された「ｘ２＞２」なる入力属性条件、または、２回目に抽出された「ｘ１＞２」なる入力属性条件、に該当するデータに「＊」を付したものである。
【００９６】
【表３３】

第2データ群DA2

id	x1	x2	x3	x4	y
5	*3	1	3	10	2
6	*4	2	3	30	2
7	1	*3	1	10	2
8	*4	*4	4	20	2

【００９７】
表３３から、問題事象（元の第２データ群ＤＡ２）に対する入力属性条件「ｘ１＞２」、「ｘ２＞２」の寄与率が表３４に示すように求められる。
【００９８】
【表３４】

問題事象に対する要因寄与率

	x1＞2	x2＞2	Total
x1＞2	50%	25%	75%
x2＞2	25%	25%	50%

【００９９】

表３４において、「ｘ１＞２」と「ｘ１＞２」との交差部に示す寄与率、及び「ｘ２＞２」と「ｘ２＞２」との交差部に示す寄与率は、それぞれ「ｘ１＞２」単独要因の寄与率、及び「ｘ２＞２」単独要因の寄与率を、それぞれ表している。また、「ｘ１＞２」と「ｘ２＞２」との交差部に示す寄与率は何れも、「ｘ１＞２」要因と「ｘ２＞２」要因との複合要因の寄与率を表している。なお、表３４は、図７のようにも表現できる。

【０１００】

表３４または図７から、問題事象（ｙ＝２）に対し、優先順位（順位１：ｘ１，順位２：ｘ２）を付けて対策を施す事ができる。

[ステップ１３]

以上でデータ分析を終了し、入力属性閾値テーブル作成部１２で作成された入力属性閾値テーブルや、寄与率のデータが、分析結果データとしてハードディスク等の分析結果データ格納部１４に格納される。この分析結果データは、適宜、分析結果データ格納部１４から表示装置や印刷装置等の出力部１５に送られ、表示装置にて決定木やテーブルとして表示したり、印刷装置にて決定木やテーブルとして印刷したりすることができる。

【０１０１】

本実施形態によれば、特許文献１に記載の、従来の決定木－２（図１２）のように、ラベル階層構造（図１１）を予め定義しなくても、表３２（または表３１）の入力属性閾値テーブルに示したような非常に簡潔な形で、問題事象の要因を導き出せる。そして、これを用いて、問題事象に対する各要因（入力属性）の寄与率を求める事ができる。

【０１０２】

ここで、表３２（または表３１）に示される本実施形態の入力属性閾値テーブルを、決定木の形式で表現すると、図８のように表される。また、従来の決定木－２（図１２）を用いて、図７と同じ形式で、問題事象ｙ＝Ｙ（＝２）に対する各要因の寄与率を表現すると、図９のようになる。

【０１０３】

本実施形態から導かれる決定木（図８）と、従来の決定木－２（図１２）とを比較すると、本実施形態の場合には、入力属性ｘ３の寄与が表現されていない。これは、図７と図９とを比較して分かるように、問題事象ｙ＝Ｙ（ｙ＝２）が、入力属性ｘ１およびｘ３の、それぞれの単独要因では発生していないからであり、上記の２回目の繰り返し操作中のステップ９において、ｘ１＞２のデータ群に対してステップ１０を実行しなかった事に因る。

【０１０４】

詳細に要因を追求する場合には、入力属性ｘ３の寄与も抽出する必要があるが、問題事象ｙ＝Ｙ（ｙ＝２）を除去する（改善する）事を目的すれば、入力属性ｘ１のみの抽出であってもこの目的を十分に達成できる。本実施形態では、この点に着目し、問題事象に対して対策すべき主要因を抽出しているため、入力属性ｘ３を抽出していない。詳細な分析を必要とする場合には、上記ステップ９で２分化されたデータ群の双方に対して、ステップ１０を実行すればよい。

【０１０５】

なお、上述した実施形態では、複数の要因を導き出し決定木を生成していたが、単に一つの要因だけを抽出したい場合であれば、ステップ８で終了してもよい。

【０１０６】

以上で説明したデータ分析方法は、コンピュータが図２のＳ０～Ｓ１２（ステップ０～１３）に対応するプロセスを含むデータ分析プログラムを実行することによって実現できる。したがって、図１のデータ分析装置は、データ分析プログラムが、コンピュータを文字－数値データ変換部１、分析対象データ格納部２、

閾値設定部３、データ分類部４、データ列抽出部５、頻度演算部６、頻度累積差演算部７、入力属性閾値決定部８、頻度累積比率演算部１６、要因抽出部９、要因未発見データ抽出部１０、終了条件判定部１１、入力属性閾値テーブル作成部１２、および寄与率演算部１３として機能させることにより実現することが可能である。

【０１０７】

上記プログラムは、コンピュータで読み取り可能な記録媒体に格納してユーザに提供することができる。この記録媒体は、コンピュータ本体に内蔵された内蔵メディアであってもよいし、コンピュータ本体に対して分離可能に構成されたリムーバブル・メディアであってもよい。上記内蔵メディアとしては、ＲＯＭ；フラッシュメモリ等の書き換え可能な不揮発性メモリ；ハードディスク等が挙げられる。また、上記リムーバブル・メディアとしては、ＣＤ－ＲＯＭ、ＤＶＤ等の光記録媒体；ＭＯ等の光磁気記録媒体；フロッピー（登録商標）ディスク、カセットテープ、リムーバブル・ハードディスク等の磁気記録媒体；メモリカード等のような書き換え可能な不揮発性メモリを内蔵したメディア；ＲＯＭカセット等のようなＲＯＭを内蔵したメディア等が挙げられる。

【０１０８】

上記プログラムは、ＣＰＵのアクセスにより実行される構成であってもよいし、記録媒体に格納されているプログラムを読み出し、読み出したプログラムを内蔵メディアのプログラム記憶領域に転送した後、内蔵メディア上のプログラムがＣＰＵのアクセスにより実行される構成であってもよい。また、上記プログラムは、コンピュータで読み取り可能な記録媒体に格納された状態で販売されるものに限定されるものではなく、インターネット等の通信ネットワークを介してユーザのコンピュータに転送する形式で販売されるものであってもよい。

【０１０９】

なお、本実施形態では、データ分類部４において出力属性と出力属性閾値との比較により分類を行っていたが、出力属性が文字属性である場合、文字－数値データ変換部１で出力属性を数値属性に変換せず、データ分類部４において出力属性と要因分析対象となる出力属性（文字；Ｙ）との比較により分類を行うようにしてもよい。

【０１１０】

本実施形態に係るデータ分析方法は、以上のように、Ｎ個（Ｎは２以上の整数）の属性からなるＮ列の入力属性のデータと、１個の属性からなる１列の出力属性のデータとで構成される基本データ群を分析対象とし、該出力属性と該入力属性との因果関係を分析するデータ分析方法であって、出力属性閾値を設定する第１ステップと、該出力属性の値と該出力属性閾値との比較に基づいて、該基本データ群を、第１データ群と第２データ群とに２分化する第２ステップと、該第１データ群および該第２データ群の各々から、第Ｊ入力属性（Ｊは、１≦Ｊ≦Ｎなる関係にある整数）のデータ列を表す１－Ｊデータ列および２－Ｊデータ列を、それぞれ抽出する第３ステップと、該１－Ｊデータ列の該第Ｊ入力属性の個々の値に対して、その値以下のデータ個数の割合を表す１－Ｊ頻度累積（％）を計算し、該２－Ｊデータ列の該第Ｊ入力属性の個々の値に対して、その値以下のデータ個数の割合を表す２－Ｊ頻度累積（％）を計算する第４ステップと、該１－Ｊデータ列および該２－Ｊデータ列の双方を含めた、該第Ｊ入力属性の全ての値の個々に対して、該１－Ｊ頻度累積（％）と該２－Ｊ頻度累積（％）との差の絶対値を表す、第Ｊ頻度累積差を計算する第５ステップと、第Ｊ頻度累積差の値が最大となるときの第Ｊ入力属性の値を第Ｊ入力属性閾値として抽出する第６ステップと、第Ｊ入力属性が第Ｊ入力属性閾値であるときにおいて、該１－Ｊ頻度累積（％）に対する該２－Ｊ頻度累積（％）の比率を表す２－Ｊ下比率、および、１００から該１－Ｊ頻度累積（％）を引いた値に対する、１００から該２－Ｊ頻度

累積（％）を引いた値の比率を表す２－Ｊ上比率を計算し、双方の比率のうちの大きい方の値を示す、２－Ｊ比率を抽出する第７ステップと、Ｊの値を１からＮまで順次増加させて、該第３ステップ～該第７ステップの操作を繰り返し、繰り返し操作中の該第７ステップで抽出された、第１から第Ｎまでの入力属性の該２－Ｊ比率のうち、その値が最大となる入力属性、該入力属性の値を表す入力属性閾値、および採用した比率の種別を抽出し、保存する第８ステップと、該第８ステップで抽出された入力属性に基づいて、該基本データ群を２分化する第９ステップと、該第９ステップで２分化されたデータ群のうちの少なくとも一方を、新たな基本データ群として、所定の終了条件を満たすまで、該第２ステップ～該第９ステップの操作を繰返す第１０ステップとを含む。

【０１１１】

上記方法によれば、ラベル階層構造を予め定義しなくても、非常に簡潔な形で問題事象の要因を複数導き出せる。そして、これを用いて、因果関係を表す決定木を作成したり、問題事象（出力属性）に対する各要因（入力属性）の寄与率を求めたりする事ができる。

なお、本発明に係るデータ分析装置は、上記の課題を解決するために、複数の入力属性と、出力属性とで構成されるデータの集合である基本データ群を分析対象とし、入力属性と出力属性との因果関係を分析し、因果関係を示す情報を抽出するデータ分析装置であって、基本データ群を出力属性に依って第１データ群と第２データ群とに分類する分類手段と、各入力属性の全ての数値について、入力属性がその数値以下であるデータが第１データ群および第２データ群のうちの一方に偏っている度合いを表す閾値評価指標を演算する第１の評価手段と、第１の評価手段で演算された閾値評価指標に基づいて、各入力属性について最大の閾値評価指標を持つ数値を各入力属性の閾値として決定する閾値決定手段と、閾値決定手段で決定された各入力属性の閾値に基づいて、「入力属性が閾値以下であれば第２データ群に含まれるデータである」という相関ルールの確からしさを表す第１のルール評価値と、「入力属性が閾値を超えていれば第２データ群に含まれるデータである」という相関ルールの確からしさを表す第２のルール評価値とを各入力属性について演算する第２の評価手段と、全ての入力属性に関する相関ルールのうちで最も高いルール評価値を持つ相関ルールの入力属性条件を示すデータを、第２データ群に対応する出力属性条件の要因を示す情報として抽出する要因抽出手段とを含むようにしてもよい。

また、本発明に係るデータ分析方法は、上記の課題を解決するために、前記のデータ分析装置を用いて、複数の入力属性と、出力属性とで構成されるデータの集合である基本データ群を分析対象とし、入力属性と出力属性との因果関係を分析し、因果関係を示す情報を抽出するデータ分析方法であって、上記分類手段により、基本データ群を出力属性に依って第１データ群と第２データ群とに分類する分類ステップと、上記第１の評価手段により、各入力属性の全ての数値について、入力属性がその数値以下であるデータが第１データ群および第２データ群のうちの一方に偏っている度合いを表す閾値評価指標を演算する第１の評価ステップと、上記閾値決定手段により、第１の評価ステップで演算された閾値評価指標に基づいて、各入力属性について最大の閾値評価指標を持つ数値を各入力属性の閾値として決定する閾値決定ステップと、上記第２の評価手段により、閾値決定ステップで決定された各入力属性の閾値に基づいて、「入力属性が閾値以下であれば第２データ群に含まれるデータである」という相関ルールの確からしさを表す第１のルール評価値と、「入力属性が閾値を超えていれば第２データ群に含まれるデータである」という相関ルールの確からしさを表す第２のルール評価値とを各入力属性について演算する第２の評価ステップと、上記要因抽出手段により、全ての入力属性に関する相関ルールのうちで最も高いルール評価値を持つ相関ルールの入力属性条件を示すデータを、第２データ群に対応する出力属性条件の要因を示す情報として抽出する要因抽出ステップとを含むようにしてもよい。

また、本発明に係るデータ分析プログラムは、上記の課題を解決するために、コンピュータを、基本データ群を出力属性に依って第１データ群と第２データ群とに分類する分類手段、各入力属性の全ての数値について、入力属性がその数値以下であるデータが第１デ

ータ群および第２データ群のうちの一方に偏っている度合いを表す閾値評価指標を演算する第１の評価手段、第１の評価手段で演算された閾値評価指標に基づいて、各入力属性について最大の閾値評価指標を持つ数値を各
入力属性の閾値として決定する閾値決定手段、閾値決定手段で決定された各入力属性の閾値に基づいて、「入力属性が閾値以下であれば第２データ群に含まれるデータである」という相関ルールの確からしさを表す第１のルール評価値と、「入力属性が閾値を超えていれば第２データ群に含まれるデータである」という相関ルールの確からしさを表す第２のルール評価値とを各入力属性について演算する第２の評価手段、および全ての入力属性に関する相関ルールのうちで最も高いルール評価値を持つ相関ルールの入力属性条件を示すデータを、第２データ群に対応する出力属性条件の要因を示す情報として抽出する要因抽出手段として機能させるためのデータ分析プログラムであってもよい。

また、本発明に係るデータ分析装置は、上記要因抽出手段で抽出された入力属性条件に基づいて、基本データ群を、上記入力属性条件を満たす要因データ群と上記入力属性条件を満たさない他データ群とに分割し、分類されたデータ群のうちの少なくとも一方を新たな基本データ群として分類手段に送る分割手段をさらに含み、分類手段による処理、第１の評価手段による処理、閾値決定手段による処理、第２の評価手段による処理、要因抽出手段による処理、および分割手段による処理からなる一連の処理が繰り返し実行されるようになっていてもよい。

【図面の簡単な説明】

【０１１２】

【図１】本発明の一実施形態に係るデータ分析装置の構成を示すブロック図である。

【図２】本発明の一実施形態に係るデータ分析方法を示すフローチャートである。

【図３】本発明の一実施形態に係るデータ分析装置における頻度累積差演算部７（ステップ５）の出力の一例をグラフで表したもので、入力属性ｘ１と、良品の１－ｘ１頻度累積（Ａ）、不良品の２－ｘ１頻度累積（Ｂ）、ｘ１頻度累積差｜Ａ－Ｂ｜との関係を示す。

【図４】本発明の一実施形態に係るデータ分析装置における頻度累積差演算部７（ステップ５）の出力の一例をグラフで表したもので、入力属性ｘ２と、良品の１－ｘ２頻度累積（Ａ）、不良品の２－ｘ２頻度累積（Ｂ）、ｘ２頻度累積差｜Ａ－Ｂ｜との関係を示す。

【図５】本発明の一実施形態に係るデータ分析装置における頻度累積差演算部７（ステップ５）の出力の一例をグラフで表したもので、入力属性ｘ３と、良品の１－ｘ３頻度累積（Ａ）、不良品の２－ｘ３頻度累積（Ｂ）、ｘ３頻度累積差｜Ａ－Ｂ｜との関係を示す。

【図６】本発明の一実施形態に係るデータ分析装置における頻度累積差演算部７（ステップ５）の出力の一例をグラフで表したもので、入力属性ｘ４と、良品の１－ｘ４頻度累積（Ａ）、不良品の２－ｘ４頻度累積（Ｂ）、ｘ４頻度累積差｜Ａ－Ｂ｜との関係を示す。

【図７】本発明の一実施形態に係るデータ分析装置における寄与率演算部１３（ステップ１２）で出力されるデータの一例であり、問題事象である出力属性条件ｙ＝２（＝Ｙ）に対する入力属性条件「ｘ１＞２」および入力属性条件「ｘ２＞２」の寄与率を示す。

【図８】本発明の実施形態の入力属性閾値テーブルを、決定木の形式で表現した図である。

【図９】従来の決定木－２を、図７と同じ形式で表現した図である。

【図１０】従来の決定木－１を表す図である。

【図１１】従来の決定木－２のラベル階層構造を表す図であり、（ａ）はｘ１属性、（ｂ）はｘ２属性、（ｃ）はｘ３属性、（ｄ）はｘ４属性を示す。

【図１２】従来の決定木－２を表す図である。

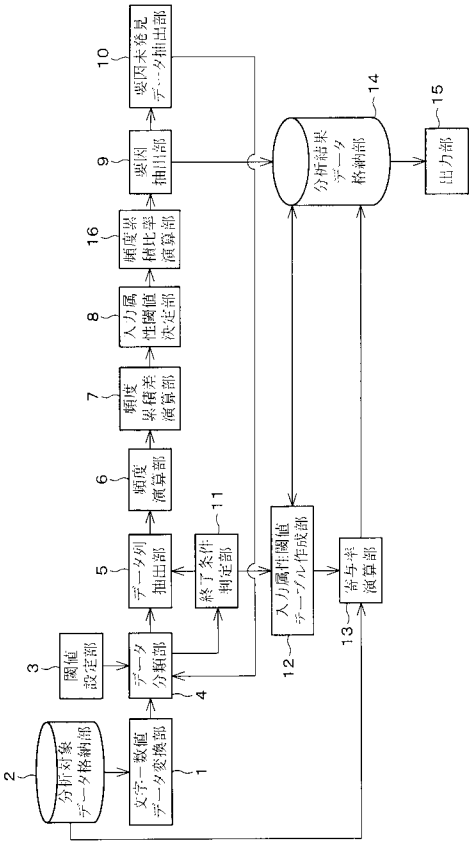
【符号の説明】

【０１１３】

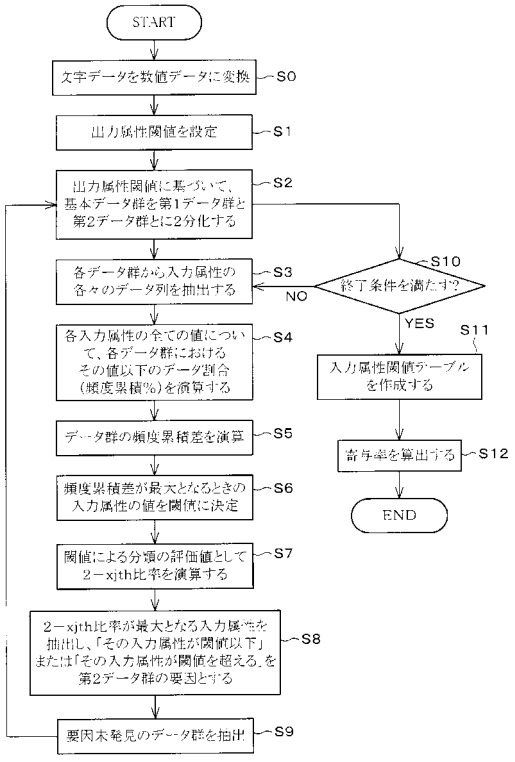
３　閾値設定部（閾値設定手段）
４　データ分類部（分類手段）
６　頻度演算部（第１の評価手段、頻度演算手段）
７　頻度累積差演算部（第１の評価手段、差分演算手段）

8　入力属性閾値決定部（閾値決定手段）
9　要因抽出部（要因抽出手段）
１０　要因未発見データ抽出部（分割手段）
１１　終了条件判定部（終了条件判定手段）
１６　頻度累積比率演算部（第２の評価手段）

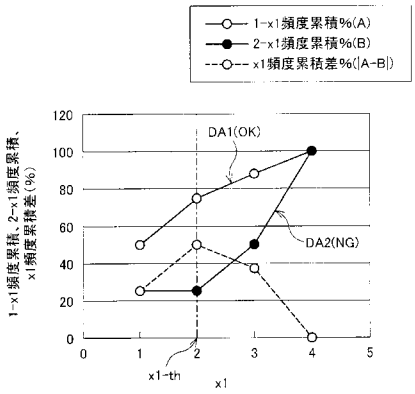
【図１】



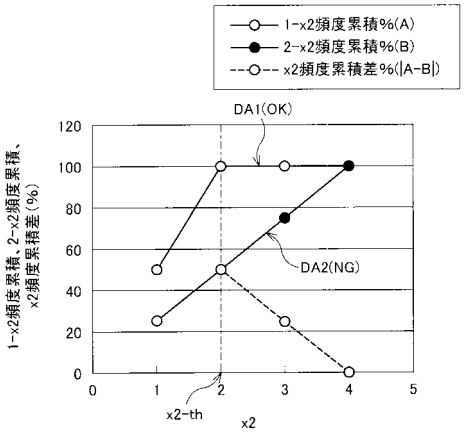
【図２】



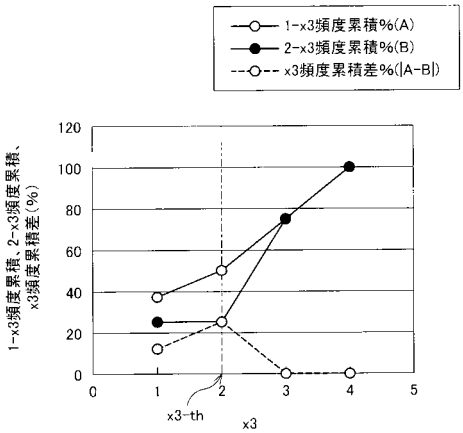
【図３】


1-x1頻度累積%(A)
2-x1頻度累積%(B)
x1頻度累積差%(|A-B|)
1-x1頻度累積、2-x1頻度累積、x1頻度累積差(%)
DA1(OK)
DA2(NG)
x1-th
x1

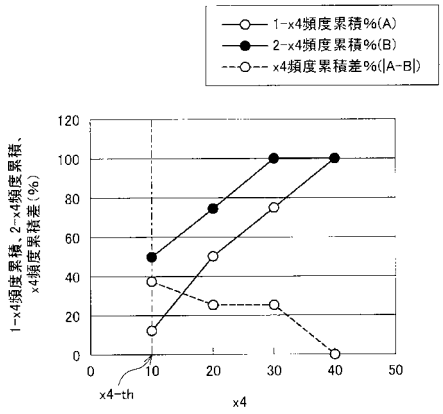
【図４】


1-x2頻度累積%(A)
2-x2頻度累積%(B)
x2頻度累積差%(|A-B|)
1-x2頻度累積、2-x2頻度累積、x2頻度累積差(%)
DA1(OK)
DA2(NG)
x2-th
x2

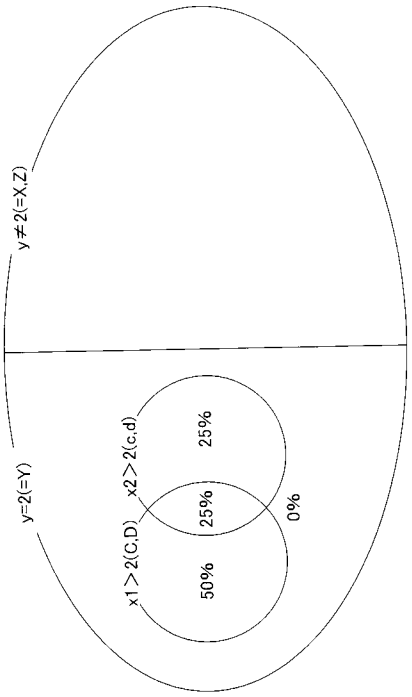
【図５】


1-x3頻度累積%(A)
2-x3頻度累積%(B)
x3頻度累積差%(|A-B|)
1-x3頻度累積、2-x3頻度累積、x3頻度累積差(%)
x3-th
x3

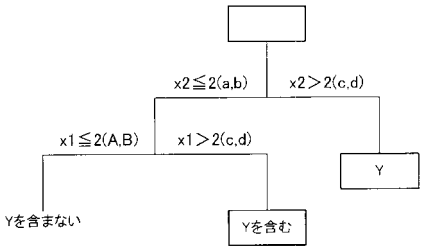
【図６】


1-x4頻度累積%(A)
2-x4頻度累積%(B)
x4頻度累積差%(|A-B|)
1-x4頻度累積、2-x4頻度累積、x4頻度累積差(%)
x4-th
x4

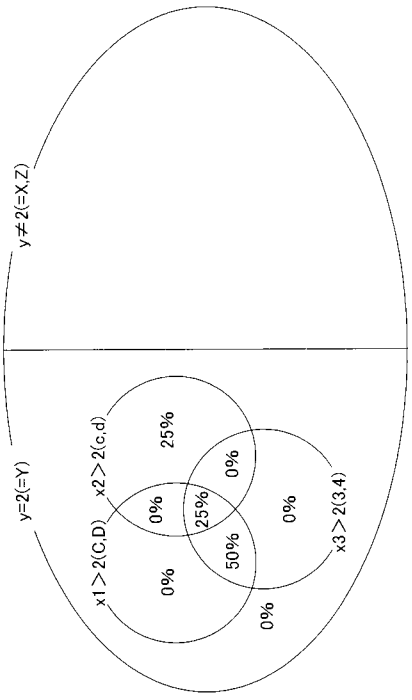
【図７】



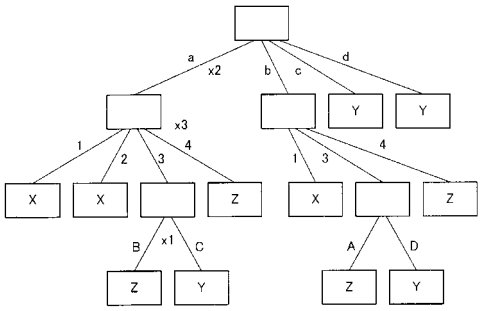
【図８】



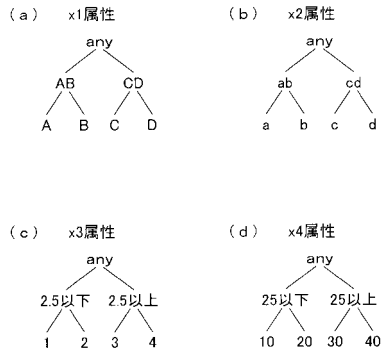
【図９】



【図１０】



【図１１】



【図１２】