

(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
2010年2月11日(11.02.2010)

(10) 国際公開番号
WO 2010/016109 A1

PCT

- (51) 国際特許分類:
G06N 5/04 (2006.01) G06Q 10/00 (2006.01)
C40B 30/02 (2006.01)
- (21) 国際出願番号: PCT/JP2008/064058
- (22) 国際出願日: 2008年8月5日(05.08.2008)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (71) 出願人(米国を除く全ての指定国について): 富士通株式会社(FUJITSU LIMITED) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 Kanagawa (JP).
- (72) 発明者; および
- (75) 発明者/出願人(米国についてのみ): 湯田 浩太郎(YUTA, Kohtarou) [JP/JP]; 〒2118588 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内 Kanagawa (JP).
- (74) 代理人: 青木 篤, 外(AOKI, Atsushi et al.); 〒1058423 東京都港区虎ノ門三丁目5番1号 虎ノ門37森ビル青和特許法律事務所 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, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), ヨーロッパ (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO,

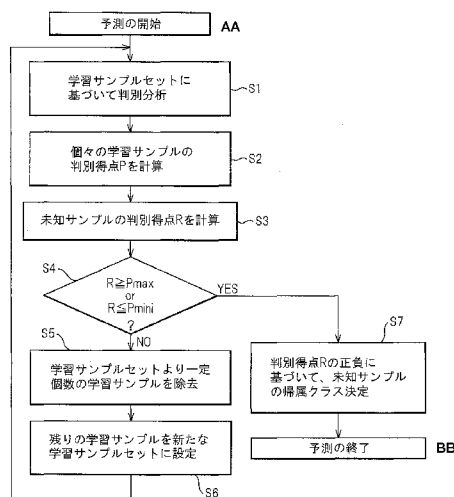
[続葉有]

(54) Title: SAMPLE ATTRIBUTION CLASS ESTIMATION METHOD, SAMPLE ATTRIBUTION CLASS ESTIMATION PROGRAM, AND ATTRIBUTION CLASS ESTIMATION APPARATUS

(54) 発明の名称: サンプルの帰属クラス予測方法、帰属クラス予測プログラムおよび帰属クラス予測装置

[xix]

Fig.3



AA START ESTIMATION
S1 DISCRIMINATE AND ANALYZE BASED ON LEARNING SAMPLE SET
S2 CALCULATE DISCRIMINATION SCORE P OF INDIVIDUAL LEARNING SAMPLES
S3 CALCULATE DISCRIMINATION SCORE R OF UNKNOWN SAMPLE
S5 ELIMINATE CONSTANT NUMBER OF LEARNING SAMPLES FROM LEARNING SAMPLE SET
S6 SET REMAINING LEARNING SET AS NEW LEARNING SAMPLE SET
S7 DETERMINE ATTRIBUTION CLASS OF UNKNOWN SAMPLE BASED ON WHETHER DISCRIMINATION SCORE R IS POSITIVE OR NEGATIVE
BB COMPLETION OF ESTIMATION

(57) Abstract: An attribution class of an unknown sample is highly accurately estimated without using an existing estimation model. The attribution class of the unknown sample is estimated by performing the following steps of a) data-analyzing a learning sample set to evaluate a discrimination function for classifying and estimating respective learning samples into a first or a second class, b) calculating discrimination scores of each learning sample and the unknown sample based on the discrimination function, c) determining whether the discrimination score of the unknown sample is larger or equal to the maximum discrimination score or smaller or equal to the minimum discrimination score among all of the learning samples, d) determining the attribution class according to the discrimination score of the unknown sample when the determination is affirmative in the step c), e) eliminating at least the learning sample having the maximum discrimination score and the learning sample having the minimum score from the learning sample set to structure a new learning sample set of the remaining learning samples when the determination is negative in the step c), and f) repeating the steps of a) to e).

(57) 要約:

[続葉有]

WO 2010/016109 A1



PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, 添付公開書類:

CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

— 国際調査報告 (条約第 21 条(3))

既成の予測モデルを使用することなく、高い精度で未知サンプルの帰属クラスを予測する。未知サンプルの帰属クラスを予測するために、a) 学習サンプルセットをデータ解析して各学習サンプルを第1または第2のクラスに分類予測する判別関数を求め、b) 前記判別関数に基づいて各学習サンプルと未知サンプルの判別得点を算出し、c) 未知サンプルの判別得点が、全ての学習サンプルのうちの最大の判別得点以上あるいは最小の判別得点以下の何れかであるか否かを判定し、d) 前記c)で肯定判断の場合、前記未知サンプルの判別得点に従ってその帰属クラスを決定し、e) 前記c)で否定判断の場合、少なくとも、前記最大の判別得点を有する学習サンプルと前記最小の判別得点を有する学習サンプルとを学習サンプルセットから除去して、残りの学習サンプルによって新たな学習サンプルセットを構築し、f) 前記a)～e)のステップを繰り返す。

明 細 書

サンプルの帰属クラス予測方法、帰属クラス予測プログラムおよび帰属クラス予測装置

技術分野

[0001] 本発明は、クラス未知のサンプルについて、そのサンプルが帰属するクラスを予測するための方法、プログラムおよび装置に関する。

背景技術

[0002] 典型的なクラス分類問題として文字認識問題がある。そのデータ解析は極めて高い水準まで達しており、一般には99%以上の認識率が達成される。一方で、環境問題等で最近注目されるようになった化合物の毒性(以下、安全性)評価のための二クラス分類問題では、文字認識問題に比べてその分類率や予測率はかなり低い。通常、分類率としては80%から90%が精々であり、予測率は70%から80%が限界と言われている。その原因は、化合物における毒性発現の要因が複雑で多岐にわたること、更に化合物間の構造が多様であること等である。ところが、化合物の安全性評価は極めて重要性の高い問題であり、その分類や予測が間違った場合、特に、毒性化合物を安全な化合物と分類予測した場合に社会に与えるインパクトは大きい。この点で、化合物の安全性評価における分類予測の正確さを改善することが強く求められている。

[0003] 従って、現在、化合物のクラス分類問題において分類率を上げることが特に重要な問題であると認識され、そのために種々の努力がなされている。本発明者は、二クラス分類問題において分類率を限りなく100%に近づけることが可能な方法、即ち、「K-step・Yard・sampling・method」(以下、KY法)を提案している(特許文献1、非特許文献1参照)。

[0004] この方法では、まず、学習サンプルセットを判別分析して、クラス1あるいはクラス2に属することが明確なサンプル群と、どちらのクラスに属するかが明確でないサンプル(グレークラスサンプル)群に別ける。クラス帰属が明確なサンプル群については判別分析の結果に基づくクラスにアサインし、グレークラスサンプル群についてはこれを

新たな学習サンプルセットに設定して新たな判別分析を行う。このプロセスをグレイクラスサンプルがなくなるまで繰り返すことにより、分類率をほぼ100%とするようにしている。

[0005] KY法を実施することによって得られた複数の判別関数は、分類目的である特性が未知のサンプルについて、その帰属クラスを予測するための予測モデルとなる。この予測モデルでは分類率がほぼ100%であるため、高い予測率を期待することができる。

[0006] 最近、EUでREACH規則の運用が開始されており、その運用に伴って化合物の毒性に関して大量のデータの蓄積が進むものと考えられる。通常、予測モデルは、サンプルの予測目的である特性値、即ち、目的変数の値が既知のサンプルを集め、これを学習サンプルセットとしてデータ解析を適用することにより作成される。学習サンプルセットに含まれるサンプル数が多いほど、信頼性の高い予測モデルを作成することができる。したがって、一旦作成された予測モデルであっても、学習サンプルとして使用可能な新たなデータが蓄積されると、そのデータを加えた新たな学習サンプルセットで新たな予測モデルを作成することが望まれる。

[0007] しかしながら、そのためには定期的に予測モデルを更新する必要があり、多大の労力と経費を要する。上述したKY法によって予測モデルを作成する場合、学習サンプルセットを何段階にもわたって判別分析する必要があり、1個の学習サンプルセットから予測モデルを作成するためには、従来の方法に比べてより多くの労力と経費を要する。したがって、固定された学習サンプルセットに基づいて作成された予測モデルを使用することなく、未知サンプルの予測を行うことができれば、極めて効率的に未知サンプルの帰属を決定することができる。また、その場合、常に新しいデータを加えた学習サンプルセットに基づいてクラス帰属を予測することが可能となるため、予測率の向上が期待される。

[0008] 特許文献1:WO2008/059624

非特許文献1:「K-step・Yard・sampling・methodの開発とADME-T予測への適用」34回構造活性相関シンポジウム、2006年11月

発明の開示

発明が解決しようとする課題

[0009] したがって、本発明の課題は、高い予測率を維持しながら、既成の予測モデルを使用することなく、常に新しい学習サンプルデータを用いて未知サンプルのクラス分類予測を行うことが可能な、サンプルの帰属クラス予測方法、そのためのプログラムおよび装置を提供することである。既成の予測モデルを必要としない点で、この方法をモデルフリー予測方法と呼ぶ。

課題を解決するための手段

[0010] 前記課題を解決するために、サンプルの帰属クラス予測方法は、a) 学習サンプルセットの各学習サンプルに対して発生させた複数の説明変数に基づいて、前記各学習サンプルを第1または第2のクラスに分類予測する判別関数を求め、b) 前記判別関数に基づいて前記各学習サンプルと未知サンプルの判別得点を算出し、c) 前記未知サンプルの判別得点が、前記全ての学習サンプルのうちの最大の判別得点以上あるいは最小の判別得点以下の何れかであるか否かを判定し、d) 前記c) で肯定判断の場合、前記未知サンプルの判別得点に従ってその帰属クラスを決定し、e) 前記c) で否定判断の場合、少なくとも、前記最大の判別得点を有する学習サンプルと前記最小の判別得点を有する学習サンプルとを前記学習サンプルセットから除去して、残りの学習サンプルによって新たな学習サンプルセットを構築し、f) 前記新たな学習サンプルセットに対して、前記a) ～e) のステップを繰り返す、各ステップを備える。

[0011] 前記課題を解決するために、サンプルの帰属クラス予測プログラムは、a) 学習サンプルセットの各学習サンプルに対して発生させた複数の説明変数に基づいて、前記各学習サンプルを第1または第2のクラスに分類予測する判別関数を求める手順と、b) 前記判別関数に基づいて各学習サンプルと未知サンプルの判別得点を算出する手順と、c) 前記未知サンプルの判別得点が、前記全ての学習サンプルのうちの最大の判別得点以上あるいは最小の判別得点以下の何れかであるか否かを判定する手順と、d) 前記c) で肯定判断の場合、前記未知サンプルの判別得点に従ってその帰属クラスを決定する手順と、e) 前記c) で否定判断の場合、少なくとも、前記最大の判別得点を有する学習サンプルと前記最小の判別得点を有する学習サンプルとを前

記学習サンプルセットから除去して、残りの学習サンプルによって新たな学習サンプルセットを構築する手順と、f) 前記新たな学習サンプルセットに対して、前記a)～e)の手順を繰り返して実行する手順と、をコンピュータに実行させる。

[0012] 前記課題を解決するために、サンプルの帰属クラス予測装置は、クラス1あるいはクラス2に帰属することが既知の複数のサンプルで構成される学習サンプルセットをデータ解析して、前記クラス1あるいはクラス2への帰属が不明である未知サンプルの帰属クラスを予測する装置であって、前記学習サンプルセットの各学習サンプルと前記未知サンプルに対して説明変数を発生する説明変数発生装置と、前記各学習サンプルについて発生した説明変数に基づいて判別分析を行い、上記クラス1およびクラス2を判別するための判別関数を得る、判別関数作成エンジンと、前記作成された判別関数に基づいて、前記各学習サンプルと前記未知サンプルの判別得点を算出する、判別得点算出装置と、前記算出された各学習サンプルの判別得点と前記未知サンプルの判別得点を比較する、判別得点比較装置と、前記判別得点比較装置の比較結果に基づいて、前記未知サンプルの帰属クラスを決定する帰属クラス決定装置と、前記判別得点比較装置の比較結果に基づいて、前記学習サンプルセットから、少なくとも最大の判別得点を有する学習サンプルと最小の判別得点を有する学習サンプルとを取り除いて新たな学習サンプルセットを構築するサンプルセット設定装置と、前記設定された新たな学習サンプルセットを前記学習サンプルセットとして用いて前記説明変数発生装置、前記判別関数作成エンジン、前記判別得点算出装置、前記判別得点比較装置、前記帰属クラス決定装置および前記サンプルセット設定装置を繰り返し動作させる制御装置と、を備え、前記帰属クラス決定装置は、前記判別得点比較装置における比較結果が、前記未知サンプルの判別得点が前記各学習サンプルのうちの最大の判別得点以上か、あるいは最小の判別得点以下かの何れかである場合、前記未知サンプルの判別得点にしたがって前記未知サンプルの帰属クラスを決定するようにしている。

[0013] 学習サンプルセットを判別分析することにより、各学習サンプルを第1のクラスまたは第2のクラスに分類予測する判別関数を得ることができる。この判別関数に基づいて計算された各サンプルの判別得点は、分類予測の確からしさを示す指標と考えら

れる。したがって、最大の判別得点を有する学習サンプル、最小の判別得点を有する学習サンプルは、判別得点の符号に従って行われるクラス分類予測の精度がもつとも高い。

[0014] 未知サンプルの判別得点を同じ判別式を用いて計算し、もしこの判別得点が、学習サンプル中の最大の判別得点以上か、あるいは最小の判別得点以下の場合、判別得点に従った未知サンプルのクラス分類精度は、前記最大あるいは最小の判別得点を有する学習サンプルのクラス分類精度と同じかそれ以上となる。したがって、この場合、未知サンプルのクラスを判別得点の、例えば符号にしたがって決定する。

[0015] 一方、未知サンプルの判別得点が、学習サンプル中の最大の判別得点より小さくかつ最小の判別得点以上の場合、判別得点にしたがって未知サンプルをクラス決定せず、学習サンプルセットから少なくとも最大、最小の判別得点を有する学習サンプルを除去して新たな学習サンプルセットを構築する。その後、この新たな学習サンプルセットについて新たな判別分析を行う。これを、未知サンプルの判別得点が、学習サンプル中の最大の判別得点以上か、あるいは最小の判別得点より小さくなるまで続けることにより、未知サンプルを非常に高い精度でクラス分類予測することができる。

発明の効果

[0016] 本明細書で開示する方法、プログラムおよび装置によれば、学習サンプルセットが存在すればそのデータ解析と並行して未知サンプルのクラス予測を行うことができる。したがって、既成の分類予測モデルを必要としないので、学習サンプルデータの蓄積に伴って分類予測モデルを更新する必要がなくなる。また、未知サンプルのクラス分類予測は、未知サンプルの判別得点が、学習サンプルセット中の最大の判別得点よりも大かあるいは最小の判別得点よりも小さいかに基づいて行われるため、その分類予測精度は極めて高くなる。

図面の簡単な説明

[0017] [図1]図1は、理想的な二クラス分類の結果としてのパターン空間を示す。

[図2]図2は、通常の二クラス分類の結果としてのパターン空間を示す。

[図3]図3は、未知サンプルのクラス帰属方法を実施する基本的な処理手順を示すフ

ローチャートである。

[図4]図4は、一実施形態の処理手順を示すフローチャートの前半部分を示す図である。

[図5]図5は、図4に示す一実施形態の処理手順の後半部分を示す図である。

[図6]図6は、サンプルデータテーブルの一例を示す図である。

[図7]図7は、最終パラメータセットのデータテーブルの一例を示す図である。

[図8]図8は、図4に示す1ステップの詳細を示す図である。

[図9]図9は、一実施形態に係る特性予測装置の概略構成を示すブロック図である。

符号の説明

- [0018]
- 1 クラス1の誤分類サンプル
 - 2 クラス2の誤分類サンプル
 - 100 サンプルの帰属クラス予測装置
 - 110 入力装置
 - 120 出力装置
 - 130 入力データテーブル
 - 140 初期パラメータセットテーブル
 - 150 最終パラメータセットテーブル
 - 160 予測結果保存テーブル
 - 200 解析部
 - 210 初期パラメータ発生エンジン
 - 220 制御部
 - 230 特徴抽出エンジン
 - 240 判別関数作成エンジン
 - 250 判別得点算出部
 - 260 サンプルセット設定部
 - 270 解析終了条件検出部
 - 280 判別得点比較部
 - 290 帰属クラス決定部

300 サンプルライブラリ

発明を実施するための最良の形態

[0019] 最適な実施形態を説明する前に、本発明の原理について説明する。

[0020] 図1は、判別分析による、理想的な二クラス分類の結果としてのパターン空間を示す。理想的とは、分類率が100%であることを意味する。図において、Sは識別面あるいは超平面を示し、識別面Sの右側の領域Aがクラス1のサンプルが存在する領域（以下、クラス1領域）であり、左側の領域Bがクラス2のサンプルが存在する領域（以下、クラス2領域）である。○は本来クラス1に属するサンプル、×は本来クラス2に属するサンプルを示す。

[0021] 二クラス分類が理想的に、即ち分類率100%で行われた場合は、識別面Sの両側にそれぞれクラス1のサンプル○、クラス2のサンプル×が位置し、クラス1の領域Aとクラス2の領域Bは重ならない。ところが、通常の二クラス分類では領域Aと領域Bに重なりが生じる。

[0022] 図2は、通常の二クラス分類の結果としてのパターン空間を示す。分類率が100%未満の場合、図示するように、識別面Sによって領域A、Bが完全に分類されず、重複領域Cが形成される。その結果、本来はクラス1のサンプルであってもクラス2領域に入り込み、クラス2サンプルとして間違って分類されるサンプル1、本来はクラス2のサンプルであってもクラス1領域に入り込み間違ってクラス1サンプルとして分類されるサンプル2が発生する。これら間違って分類されるサンプルを誤分類サンプルと呼び、正しく分類されるサンプルを正分類サンプルと呼ぶ。従って、誤分類サンプルには、本来クラス1であるサンプルおよび本来クラス2であるサンプルの両者が含まれる。

[0023] クラス1とクラス2を判別するための判別関数（予測モデル）Yは、以下の式（1）で示される。

$$[0024] \quad Y = a_1x_1 \pm a_2x_2 \pm \cdots \pm a_nx_n \pm C \quad (1)$$

[0025] 式（1）において、Yは目的変数の値を示し、 x_1 、 x_2 、 $\cdots$ 、 x_n は説明変数（以下、パラメータ）の値、 a_1 、 a_2 、 $\cdots$ 、 a_n は重み係数、Cは定数を示す。説明変数の値 x_1 、 x_2 、 $\cdots$ 、 x_n は各サンプルによって異なる値であり、重み係数 a_1 、 a_2 、 $\cdots$ 、 a_n は全サンプル

ルに共通の値である。通常、クラス1に帰属するサンプルは目的変数 Y の値が正の値に、クラス2に属するサンプルは目的変数 Y の値が負の値になるように重み係数が選択されている。この場合、図1および2に示す識別面 S は、 Y の値が0となる平面(超平面)を示す。

[0026] 式(1)に、各サンプルについて求めた説明変数(パラメータ)の値 $x_1, x_2, \dots, x_n$ を代入することにより、各サンプルの目的変数の値 $Y(k)$ を求めることができる。この値 $Y(k)$ は、判別分析においては「目的変数」あるいは「独立変数」として意味づけされるが、数値データとしてみた場合、連続変数である「判別得点」を示す。判別得点は、パターン空間上において、各サンプルと識別面 S との間の距離を示している。

[0027] ニクラス分類において、通常、クラス1サンプルの判別得点はプラス(正)の符号を有し、クラス2サンプルの判別得点はマイナス(負)の符号を有するように判別関数が作成されている。従って、サンプルの帰属クラスの決定には、判別得点がプラスであるか、マイナスであるかしか意味を持たない。判別得点の大きさ(絶対値)が、分類目的の特性の大きさあるいは強さを示すものでは決してない。しかしながら、判別得点 $Y(k)$ は、各サンプルによって特定されるパラメータ値を式(1)に当てはめて計算するものであるため、判別得点 $Y(k)$ に基づく各サンプル間の関係は固定されている。

[0028] 図2に示すように、誤分類サンプル1、2は、識別面 S の近傍において発生する傾向がある。即ち、識別面 S とサンプル間の距離が小さい領域において誤分類サンプル1、2が発生しやすい。識別面 S とサンプル間の距離が大きくなる程、誤分類サンプルの発生の度合いが小さくなる。判別得点 $Y(k)$ は、識別面 S とサンプル間の距離を表すもので、従って、判別得点が高いほど、あるいは小さいほど、誤分類サンプルの発生確率が低くなる。即ち、判別得点は、クラス帰属の確からしさを示す指標であると考えてもよい。

[0029] 一方、目的変数の値が未知のサンプルについて、目的変数の予測値を求める場合、即ち帰属クラスを予測する場合、未知サンプルについて説明変数の値を算出し、これを式(1)に代入して Y の値を決定する。値 Y の符号にしたがって、未知サンプルが所属するクラスを決定する。例えば、 Y の値がプラスの場合、未知サンプルをクラス1にアサインし、 Y の値がマイナスの場合、未知サンプルをクラス2にアサインする。

- [0030] この場合、図2を参照することによって明らかなように、未知サンプルの判別得点 Y の絶対値が大きい場合、パターン空間上での未知サンプルの位置10aが識別面 S から充分はなれていると考えられる。したがってその場合の予測結果は正しいと考えられる。一方、未知サンプルの判別得点 Y の絶対値が小さい場合は、パターン空間上での未知サンプルの位置10bは識別面 S からあまり離れていないと考えられる。そのため、未知サンプルの位置10bが誤分類領域 C にある可能性が高く、予測結果の信頼性は低い。
- [0031] したがって、未知サンプル10a、10bの予測が正確か否かを判定するためには、誤分類領域 C を特定することが必要である。本発明者は、KY法の一手法として、誤分類領域 C を各学習サンプルの判別得点に基づいて決定し、その結果に対して新たな判別分析を行うことにより、ほぼ100%の分類率を有する予測モデルを作成する方法を既に提案している(PCT/2007/074334)。
- [0032] この方法では、学習サンプルセットに対して求めた判別関数に基づいて個々の学習サンプルの判別得点を計算し、個々の学習サンプルの帰属クラスを予測する。この予測クラスと実測値に基づくクラスとを比較し、両者が一致しない場合そのサンプルを誤分類サンプルとする。誤分類サンプル中で最大、最小の判別得点を検出し、その間の領域を誤分類領域 C としている。誤分類領域 C が特定されれば、判別得点に基づく未知サンプル10a、10bの予測が正しいか否かがわかる。
- [0033] この方法では、1個の判別関数を求めた時点で、学習サンプル個々の判別得点を求めると同時に、その値を実測値と比較して個々の学習サンプルが誤分類サンプルか正分類サンプルかを特定する必要がある。また、判別手法によって誤分類サンプルの発生の度合いが変化するため、誤分類領域 C を正確に特定することはできない。このような観点から、本発明者は、誤分類領域 C を特定することなく分類予測を実行することができれば、予測プロセスが簡単となり、さらに予測精度も向上すると考えた。
- [0034] 誤分類サンプルは識別面 S の近傍に存在する確率が高い。一方、判別得点が識別面 S ($Y=0$)から離れるほど、サンプルの分類予測は正確となる。図2から明らかなように、最大の判別得点 $P_{\max}$ を有するサンプル、最小の判別得点 $P_{\min}$ を有するサ

ンプルは正分類サンプルである。また、これらのサンプルの近傍のサンプルも正分類サンプルである。あるいは正分類サンプルである確率が非常に高い。したがって、判別得点に基づいて未知サンプルのクラス予測を行う場合、未知サンプルの判別得点 R が、学習サンプル中の最大の判別得点 $P_{\max}$ 以上か(点 R_a)、最小の判別得点 $P_{\min}$ 以下の場合(点 R_b)、その未知サンプルは確実にクラス1サンプル、あるいはクラス2サンプルであると予測することができる。

[0035] したがって、未知サンプルの判別得点 R が学習サンプル中の最大の判別得点 $P_{\max}$ 以上か、あるいは最小の判別得点 $P_{\min}$ 以下の場合、判別得点に基づくクラス予測を正確であると判断し、それ以外の場合、クラス予測は正確でないと判断することにより、誤分類領域 C を特定することなく、未知サンプルのクラス予測を高い精度で実施することができる。

[0036] 図3は、以上に述べたような予測の正確性に関する判断基準に基づいて、モデルフリーで未知サンプルのクラス帰属予測を行う基本的手順を示すフローチャートである。なお、図3では一個の未知サンプルのクラス帰属をモデルフリーで予測する手順を示しているが、このフローチャートを同時に複数の未知サンプルのクラス帰属を予測する場合に適用させることは容易である。

[0037] まず、ステップS1において、目的変数の実測値、即ち、クラス1であるかクラス2であるかが既知の学習サンプルを複数個用意し、判別分析用のシステムに入力して、これらについて二クラス判別分析を行う。次のステップS2では、ステップS1の判別分析によって求めた判別関数を用いて、個々の学習サンプルの判別得点を計算する。ステップS3では、同じ判別関数を用いて目的変数が未知のサンプルの判別得点を計算する。ステップS4では、ステップS2及びS3で求めた判別得点を大小順にソートし、未知サンプルの判別得点 R と、学習サンプルの最大の判別得点 $P_{\max}$ および最小の判別得点 $P_{\min}$ とを比較する。この比較の結果、未知サンプルの判別得点 R が学習サンプル中の最大の判別得点 $P_{\max}$ 以上か、あるいは最小の判別得点 $P_{\min}$ 以下と判断されると(ステップS4のYES)、ステップS7において、未知サンプルの判別得点 R の正負に基づいて未知サンプルのクラスアサインを行い、予測を終了する。

[0038] 一方、ステップS4で、未知サンプルの判別得点 R が、学習サンプル中の最大の判

別得点 $P_{\max}$ よりも小さくかつ最小の判別得点 $P_{\min}$ よりも大きいと判定される(ステップS4のNO)と、ステップS5に移行して、学習サンプルセットからある数の学習サンプルを除去し、新たなサンプルセットを構築する。

- [0039] 除去する学習サンプルは、1)判別得点の最大のサンプルから判別得点の大きい順に m 個と、判別得点の最小のサンプルから判別得点の小さい順に n 個のサンプルを指定するようにしても良い。この場合、 $m=n$ であっても良いし、 $m \neq n$ であっても良い。また、クラスポピュレーションを考慮に入れて m と n の値を決定しても良い。あるいは、2)最大あるいは最小の判別得点を100%とした場合、例えば100%から90%の範囲内の判別得点を有するサンプルを除去するようにしても良い。個数 m 、 n があまり大きいと、取り除くサンプルの範囲が、最大、最小の判別得点を有するサンプルの近傍から外れ、誤分類領域に入り込む。したがって、学習サンプルの総数の10%以下、数%程度が望ましい。
- [0040] 最も精度の高い分類予測を行おうとする場合には、ステップS4で除去するサンプルは、判別得点が最大のサンプルと、判別得点が最小のサンプルの計2個である。したがって、ステップS4では、少なくとも、判別得点が最大の学習サンプルと、判別得点が最小の学習サンプルとを取り除く。なお、図3において、後述するステップS1からステップS6までの繰り返しの最大の回数をシステムにおいて設定した場合、この最大回数に基づいて、1回の判別分析で取り除くサンプル数を決定するようにしても良い。例えば、総サンプル数が1000個であり、繰り返しの最大回数を100回とした場合、1回の判別分析では10個のサンプルを取り除くようにすれば、100回の判別分析で全てのサンプルを取り除くことができる。
- [0041] 図3のステップS6では、残りの学習サンプルによって新たな学習サンプルセットを構築し、ステップS1に戻って新たな判別分析を開始する。この場合、学習サンプルセットが初期学習サンプルセットとは異なっているので、特徴抽出の結果、新たな説明変数の組合せ(最終パラメータセット)が選択され、第1回目の判別分析とは異なる判別関数が得られる。したがって、第1回目の判別分析によるパターン空間と、第2回目の判別分析によるパターン空間とは異なったものとなる。また、個々の学習サンプルおよび未知サンプルの判別得点も第1回目の判別分析における判別得点とは異

なってくる。

- [0042] したがって、ステップS4では、学習サンプルと未知サンプルについて新に算出された判別得点に基づいて、その大小が比較される。ステップS4でYESと判定されると、ステップS7で未知サンプルの判別得点の符号に基づいてそのクラス帰属が決定される。ステップS4でNOと判定されると、ステップS5以降を実行する。
- [0043] 以上のように、ステップS1～ステップS6をステップS4でYESとなるまで繰り返すことにより、誤分類領域(グレーゾーン)を特定することなく、未知サンプルのクラス帰属を正確に予測することができる。なお、システムにおいて、繰り返し数の最大値を予め設定しておいても良い。
- [0044] 従来の分類予測方法では、学習サンプル数の数が多くなればなる程、分類率、予測率は低下する。これに対して、図3の方法では、学習サンプルの数が多くなった場合であっても、ステップS1～ステップS6の繰り返しの回数を多くするだけで、クラス予測の精度を落とすことなく、未知サンプルのクラス帰属を予測することができる。したがって、REACH規則の運用によって何万、あるいは何十万という化合物データが蓄積された場合であっても、これらを全て学習サンプルとして、しかも高いに予測精度を維持して、未知サンプルのクラス帰属を予測することができる。
- [0045] 以下に、第1の実施形態について説明する。この実施形態では、サンプルを化合物とし、予測を行う目的変数をAmesテストにおける活性(+、mutagen)、非活性(－、non-mutagen)としている。
- [0046] 図4および図5は、化合物のAmesテストにおける活性、非活性を予測するための実施手順を示すフローチャートである。まず、ステップS10において、学習サンプルセットおよび予測対象化合物である未知サンプルを用意し、システムに入力する。学習サンプルセットは、Amesテストにおける活性、非活性の実測値が既知の複数の化合物で構成される。未知サンプルは、化学的構造は既知であるがAmesテストにおける活性、非活性の実測値が不明のサンプルである。未知サンプルは1個であっても、複数個を用意しても何れでも良い。
- [0047] 用意された学習サンプルは、1次元、2次元あるいは3次元構造式の形でサンプルの帰属クラス予測装置に入力され、図6に示すようなサンプルデータを保存するため

のテーブルが作成される。図6において、コラム60はサンプルである化合物の2次元あるいは3次元の構造式を示す。コラム61はその化合物のCAS番号を示し、コラム62は、Amesテストの結果を示している。コラム62において、mutagenはAmesテストの結果、変異原性有り(+)を示し、nonmutagenは変異原性を持たない(-)ことを示している。図示の例では、mutagenであるサンプルをクラス1(ポジティブクラス)に、nonmutagenであるサンプルをクラス2(ネガティブクラス)に二クラス分類するためのデータテーブルを示している。コラム63は、サンプル番号を示す。未知サンプルについても、同様のデータテーブルが作成されるが、Amesテストの実測値を記載するコラム62は空欄である。

[0048] 次に、図4のステップS11において、目的変数を算出するための初期パラメータ、即ち説明変数($x_1, x_2 \dots x_n$)の値を発生させる。初期パラメータは、化合物の構造から自動的に発生させることができる。例えば、富士通株式会社で販売するADME WORKS—ModelBuilder(登録商標)では、化合物の2次元あるいは3次元構造、各種の物性に基づいて自動的に約800個のパラメータを発生させることができる。初期パラメータは、各学習サンプルおよび未知サンプルについて発生させる。ステップS12では、第1回目の判別分析を開始するために、STAGEを1に設定する。

[0049] ステップS13では、学習サンプルの初期パラメータに対して特徴抽出を行い、分類に不必要なノイズパラメータを除去する。これによって、最終パラメータセット($x_1, x_2 \dots x_n$)が決定される(ステップS14)。特徴抽出としては、単相関係数、重相関係数、出現頻度、Fischer比、Variance法などの種々の既知手法を用いて実施することができる。特徴抽出のための各種のエンジンも一般に提供されている。

[0050] 図7は、特徴抽出の結果、Amesテスト結果に影響ありとして選択された最終パラメータセットと、個々の化合物のこれらのパラメータに対する数値データを示すテーブルである。コラム70は化合物を構造式で特定し、コラム71以降は各種のパラメータを示している。例えば、コラム71は化合物の分子量を、コラム72は分子表面積を、コラム73はlogPの値をパラメータとしたことを示している。データテーブル中のセル74内に記載された値は、サンプル1分子の分子量を示すデータ、セル75内の値はサンプル1の分子表面積の値を示すデータ、セル76内の値はサンプル1のlogP値を示す

データである。各セル内に示された値が、そのサンプルのパラメータデータとなる。なお、コラム77は各サンプルのサンプル番号を示している。

- [0051] ステップS14では、解析の終了条件を満足したか否かが判定される。解析の終了条件とは、例えば、クラスアサイン(クラス予測)を行おうとする未知サンプルが存在しないこと、予め設定された最大のSTAGE数に達したか否か、学習サンプルセット中のサンプル数が予め設定した一定個数以下となったか否か、あるいは、信頼性指標が予め決定された値以下となったか否か、等である。ここで、信頼性指標とはサンプル数をパラメータ数で除した値であり、この値が小さくなると、これらのサンプルとパラメータとから得られる判別式ほとんど科学的およびデータ解析的に意味を持たず、その解析は失敗したと判断される。通常の判別分析ではこの値として4が設定されるが、4よりも大きければ大きいほど信頼性の高い判別式とみなされる。
- [0052] 図8に、ステップS15の詳細を示す。ステップS151からステップS154が図4のステップS15に相当する。まず、ステップS151で、予測を行おうとする未知サンプルが存在するか否かを確認する。未知サンプルが存在しない場合(ステップS151のYES)は、予測処理を終了する。ステップS152で、現在の処理STAGEが予め決定しておいた最大のSTAGE数N以上となったか否かを判定する。STAGE数がN以上となった場合(ステップS152のYES)は、処理を終了する。
- [0053] ステップS153では、サンプル数とパラメータ数で決定される信頼性指標の値が4未満となったか否かを判定する。信頼性指標が4以下となった場合(ステップS153のYES)は、処理を終了する。ステップS154では、学習サンプル数が予め決定した最低のサンプル数Sminiに達したか否かを判定する。最低のサンプル数以下になった場合(ステップS154のYES)は、処理を終了する。
- [0054] 従って、図示の実施形態では、終了条件の全てが満足されない場合に、ステップS16が実行される。なお、ステップS151からステップS154の実行順序は任意でよく、また、ステップS152、154等は省略しても良い。更に、ステップS152に代わって、あるいは追加のステップとして、処理時間に関する制限を入れても良い。
- [0055] 再び図4に戻ると、ステップS15で終了条件が満足される(ステップS15のYES)と、解析処理を終了する。ステップS15で終了条件に達しない場合(ステップS15のN

O)、ステップS16で最終パラメータセットを用いて第1段階(STAGE1)の判別式を構築する。判別分析では、判別関数は上記式(1)として示される。

[0056] ステップS17では、作成した判別関数を用いて各学習サンプルの判別得点を計算し、更に、同じ判別関数を用いて未知サンプルの判別得点を計算する。k番目のサンプルの判別得点 Y_k は、

$$Y_k = a_1 x_{1k} \pm a_2 x_{2k} \pm \dots \pm a_n x_{nk} \pm C \quad (2)$$

として計算される。ここで、 x_{1k} 、 x_{2k} 、 $\dots$ 、 x_{nk} は、k番目のサンプルのパラメータ(説明変数)データ、 a_1 、 a_2 、 $a_3 \dots a_n$ は各パラメータに対する重み係数であり、各サンプルについて共通である。Cは定数である。

[0057] パラメータデータ x_{11} 、 x_{21} 、 $x_{31} \dots$ は、図7の各セル内に記載されたデータから得られる。したがって、判別分析により、各パラメータに対する係数 a_1 、 $a_2 \dots$ を求めると、図7のテーブルに示される各セル内のデータを式(2)に導入することによって、サンプルkの判別得点 Y_k が計算される。

[0058] 図5に示すステップS18では、判別得点 Y_k を全学習サンプルについてソートすることにより、学習サンプルの最大の判別得点 $P_{\max}$ 、最小の判別得点 $P_{\min}$ を特定する。ステップS18では、ステップS17で求めた未知サンプルの判別得点Rと、学習サンプルの最大、最小の判別得点 $P_{\max}$ 、 $P_{\min}$ とを比較し、 $R \geq P_{\max}$ あるいは $R \leq P_{\min}$ を満足する未知サンプルがあるか否かを判定する。複数の未知サンプルのうち、 $R \geq P_{\max}$ あるいは $R \leq P_{\min}$ を満足する未知サンプルがあれば(ステップS19のYES)、そのサンプルを取り出し、判別得点の符号にしたがってクラスアサインする(ステップS20)。

[0059] 次に、再びステップS19に戻って、 $R \geq P_{\max}$ あるいは $R \leq P_{\min}$ を満足する他の未知サンプルがあるか否かを判定する。ステップS19とステップS20は、ステップS19において R (未知サンプルの判別得点) $\geq P_{\max}$ あるいは $R \leq P_{\min}$ を満足する他の未知サンプルがなくなるまで、繰り返して実行される。この結果、ステップS20において、現在のSTAGEで $R \geq P_{\max}$ あるいは $R \leq P_{\min}$ を満足する全ての未知サンプルが取りだされ、そのクラスが決定される。

[0060] ステップS19でNOとなった場合、ステップS21において、学習サンプルセットの各

クラスから判別得点の大きいあるいは小さいサンプルを、サンプル除去のルールに従って除去する。サンプル除去のルールは、図3を参照して説明したいずれのルールでも良く、ユーザが任意に指定すれば良い。ステップS22では、残りの学習サンプルを新たな学習サンプルセットに設定する。ステップS23ではSTAGEを1だけ進めてSTAGE=2とし、その後、図5のステップS13以下を実行することにより、STAGE2での未知サンプルのクラスアサイン(帰属クラス決定)を実行する。

[0061] 以上のようにして、ステップS13からステップS23を繰り返し実行することにより、全ての未知サンプルのクラスアサインが終了すると、ステップS15で解析終了条件が満足されるため、その時点で分類予測処理が終了する。この場合、未知サンプルの帰属クラス決定は、未知サンプルの判別得点が学習サンプルの最大の判別得点 P_{max} 以上となった場合、あるいは最小の判別得点 P_{mini} 以下となった場合に行われるため、その予測の信頼性は非常に高いものとなる。

[0062] なお、各STAGEで実施する判別分析手法としては、各STAGEで同じ手法を使用しても良いし、あるいは異なる手法を使用しても良い。例えば、STAGE1でBayes判別分析法を使用し、STAGE2でAdaBoost法を使用するようにしても良い。更に、各STAGEで除去する学習サンプルの個数を一定にしても良いし、あるいはSTAGE毎に異なる個数を設定しても良い。例えば、比較的初期のSTAGEでは比較的多くのサンプルを除去し、高次のSTAGEではそれよりも小さい数のサンプルを除去するようにしても良い。

[0063] 以上の結果、モデルフリーでかつ高い予測精度で、未知サンプルの帰属クラス予測を行うことができる。また、予測過程の各STAGEにおいて、各学習サンプルを正分類サンプルと誤分類サンプルとに分類する必要がなくなるので、予測のための計算時間を大幅に短縮することができる。

[0064] 図9は、第3の実施形態に係るサンプルの帰属クラス予測装置の概略構成を示すブロック図である。本実施形態にかかる予測装置100は、学習サンプル、未知サンプルのデータを入力する入力装置110、未知サンプルの予測結果あるいは処理途中の必要なデータを出力する出力装置120を備えている。サンプルが化合物の場合、入力装置110から化合物の1次元構造式あるいは2次元構造式と共に、その化合物

の既知の目的変数の値(即ち、クラス1かクラス2かに関する情報、例えば、安全性データ)が入力され、入力データテーブル130に保持される。未知サンプルを入力する場合は目的変数の値は入力されない。

- [0065] 入力装置110を介して、初期パラメータセットのデータを入力し、これを初期パラメータセットテーブル140に保存するようにしても良い。解析部200が、入力されたサンプルについて初期パラメータを自動的に発生するためのエンジン210を有している場合は、初期パラメータセットデータを入力装置110から入力する必要はない。
- [0066] 図9において、150は最終パラメータセットを保存するテーブルであり、初期パラメータセットに対して特徴抽出を行った結果としての最終パラメータセットを保存する。160は、未知サンプルについてのクラスアサイン情報、即ち予測結果情報を保存する予測結果保存テーブルである。通常、これらのテーブル130、140、150、160は一個の記憶装置に格納されている。
- [0067] 解析部200は、制御部220と、初期パラメータ発生エンジン210、特徴抽出エンジン230、判別関数作成エンジン240、判別得点算出部250、サンプルセット設定部260および解析終了条件検出部270、判別得点比較部280およびクラスアサイン実施部290を備えている。なお、初期パラメータを本装置の外部で発生させる場合は、初期パラメータ発生エンジン210は必要とされない。サンプルが化合物の場合は、初期パラメータ発生エンジン210は、入力装置110から入力された化合物の構造式に基づいて、複数の構造記述子を発生する。300はサンプルライブラリであり、過去の特性予測に用いた学習サンプルの構造式、安全性データ、初期パラメータ値などをライブラリデータとして保存している。
- [0068] 特徴抽出エンジン230は、初期パラメータセットに対して特徴抽出を行って最終パラメータセットを決定し、これを最終パラメータセットテーブル130に保存する。判別関数作成エンジン240は、種々の既存の判別分析エンジンを備えており、ユーザによって任意に指定された判別分析エンジン、あるいはシステムが適宜選択した判別分析エンジンを用いて、最終パラメータセットテーブル130を参照しながら、入力サンプルの判別分析を行い、判別関数を作成する。判別得点算出部250は、判別関数作成エンジン240によって作成された判別関数に各サンプルのパラメータを入力するこ

とにより、各サンプルの判別得点を計算する。サンプルセット設定部260は、予め設定したサンプル除去ルールに従って、学習サンプルセットから判別得点の大きいサンプル、小さいサンプルを除去し、新たな学習サンプルセットを構築する。

[0069] 特徴抽出エンジン230、判別関数作成エンジン240、判別得点算出部250、新たなサンプルセット設定部260は、制御部220の制御下で作動し、図5および図6に示す処理を実行する。なお、解析終了条件検出部270は、1) 帰属クラスを予測する未知サンプルが0となった時点、2) 処理の繰り返しが予め設定した回数(あるいは時間)を超えた場合、3) 新たに設定した学習サンプルセット中のサンプル数が予め決定した個数以下となった時点、4) 信頼性指標が予め定めた値以下となった時点、の何れかで、解析の実行を停止する。

[0070] 判別得点比較部280は、未知サンプルの判別得点Rを、学習サンプルの最大の判別得点Pmaxあるいは最小の判別得点Pminiと比較する。帰属クラス決定部290は、判別得点比較部280の結果に基づいて未知サンプルのクラスアサインを実行する。即ち、未知サンプルの判別得点Rが学習サンプルの最大の判別得点Pmax以上の場合あるいは最小の判別得点Pmini以下の場合、未知サンプルを判別得点Rの符号に基づくクラスにアサインし、その何れでもない場合、クラスアサインを行わない。帰属クラス決定部290の結果は、予測結果保存テーブル160に一旦保存された後、出力装置120を介して任意の形状で出力される。出力装置としては、種々の記憶装置、ディスプレイ、プリンタ等が選択可能であり、出力の形態としては、各種ファイル(例えばUSBファイル)、ディスプレイ表示、プリントアウト等を適宜選択することが可能である。

[0071] 上記した各プログラムは、コンピュータ読取可能な記録媒体に格納することが可能であり、これらの記録媒体を頒布、流通することが可能である。また、同様に上記した各プログラムはインターネット等の通信回線を介して頒布、流通することが可能である。前記したコンピュータ読取可能な記録媒体としては、磁気記録装置、光ディスク、光磁気ディスクあるいは半導体メモリ(RAM、ROM等)などがある。一例として、磁気記録装置としてはハードディスクドライブ(HDD)、フレキシブルディスク(FD)、磁気テープ(MT)等がある。光ディスクの例としては、DVD(Digital・Versatile・Disc)、D

VD-RAM、CD-ROM、CD-R、CD-RW等がある。光磁気ディスクとしてはMO(Magneto-Optical・disc)等がある。

[0072] 入力装置110を介して入力された学習サンプルのデータ、これらの学習サンプルについて発生された初期パラメータセットのデータは、解析部200を介してサンプルライブラリに入力され保存される。従って、サンプルライブラリ300を利用すれば、未知サンプルのクラス帰属予測を行う場合、入力装置110を介して未知サンプルの構造式データを入力するのみで、予測を実行することができる。あるいは、新たな学習サンプルデータが用意されれば、そのデータを入力装置110から帰属クラス予測装置100に入力することにより、サンプルライブラリ300に保存された既存の学習サンプルデータと合わせて最新の学習サンプルセットを作成することができる。クラス予測は、この最新の学習サンプルセットを初期サンプルセットとして用いて行われる。従って、本装置では、常に新しい学習サンプルを加えたデータベースにより、未知サンプルの特性予測を実行することができる。

[0073] また、入力装置110および出力装置120を、対話形式のグラフィックユーザインターフェースを備えたディスプレイ装置で構成することも可能である。この場合、解析終了条件の選択、サンプル除去ルールを選択などを、ユーザがディスプレイ画面を介してコンピュータと対話しながら行うようにしても良い。

産業上の利用可能性

[0074] 本明細書に開示する方法、プログラム、装置は、二クラス分類が適用可能な全ての産業分野に適用可能である。以下に、主な適用分野を列挙する。

- [0075]
- 1) 化学データ解析
 - 2) バイオ関連研究
 - 3) 蛋白質関連研究
 - 4) 医療関連研究
 - 5) 食品関連研究
 - 6) 経済関連研究
 - 7) 工学関連研究
 - 8) 生産歩留まり向上等を目的としたデータ解析

9) 環境関連研究

1) の化学データ解析分野では、より詳細には、下記のような研究に適用できる。

(1) 構造－活性／ADME／毒性／物性相関の研究

(2) 構造－スペクトル相関研究

(3) メタボノミクス関連研究

(4) ケモメトリクス研究

[0076] 例えば、構造－毒性相関研究分野では、Amesテスト結果の予測を行うことが極めて重要である。何故ならば、Amesテストは、毒性化合物規制関連の化合物審査法や労働安全衛生法等の国レベルの化合物規制に、最重要項目の一つとして組み込まれているからである。このAmesテストの審査を通らなければ、日本国内での化合物生産はできなくなり、企業の生産活動自体がストップする。また、海外での生産や輸出等も対象国の安全性規制により活動できなくなる。

[0077] ヨーロッパ議会におけるREACH規則では、化合物を使用する企業が、その化合物についてAmesテストの結果を予測し、評価する義務を負っている。本発明は、このようなテスト結果の予測に、非常に有用なツールを提供するものである。なお、Amesテストとは、米国のAmes博士が開発した変異原性試験の一つで、発がん性試験の簡易手法である。このために、多くの化学物質あるいはこれらを用いた製品の安全性の指針として採用されている試験である。

請求の範囲

- [1] a) 学習サンプルセットの各学習サンプルに対して発生させた複数の説明変数に基づいて、前記各学習サンプルを第1または第2のクラスに分類予測する判別関数を求め、
- b) 前記判別関数に基づいて前記各学習サンプルと未知サンプルの判別得点を算出し、
- c) 前記未知サンプルの判別得点が、前記全ての学習サンプルのうちの最大の判別得点以上あるいは最小の判別得点以下の何れかであるか否かを判定し、
- d) 前記c)で肯定判断の場合、前記未知サンプルの判別得点に従ってその帰属クラスを決定し、
- e) 前記c)で否定判断の場合、少なくとも、前記最大の判別得点を有する学習サンプルと前記最小の判別得点を有する学習サンプルとを前記学習サンプルセットから除去して、残りの学習サンプルによって新たな学習サンプルセットを構築し、
- f) 前記新たな学習サンプルセットに対して、前記a)～e)のステップを繰り返す、各ステップを備える、サンプルの帰属クラス予測方法。
- [2] 請求項1に記載の方法において、さらに、g) 帰属クラス未定の未知サンプルが存在するか否かを判定するステップを備え、当該ステップで否定判断の場合、前記ステップf)の繰り返しを停止する、サンプルの帰属クラス予測方法。
- [3] 請求項1に記載の方法において、前記ステップa)における判別関数は、前記複数の説明変数に特徴抽出を実施して不要な説明変数を除去した残りの説明変数に基づいて求める、サンプルの帰属クラス予測方法。
- [4] 請求項3に記載の方法において、前記ステップf)の繰り返しは、前記学習サンプルセット中の総サンプル数と、前記特徴抽出後の説明変数の数との比が予め決定した一定数以下となった場合、停止する、サンプルの帰属クラス予測方法。
- [5] 請求項1に記載の方法において、前記ステップe)では、前記最大の判別得点を有する学習サンプルから判別得点の大きい順に第1の個数の学習サンプルを、前記最小の判別得点を有する学習サンプルから判別得点の小さい順に第2の個数の学習サンプルを、前記学習サンプルセットよりそれぞれ除去する、サンプルの帰属クラス

予測方法。

- [6] 請求項5に記載の方法において、前記第1の個数と前記第2の個数は、前記学習サンプルセットにおいて、実測値が前記第1クラスである学習サンプルの数と前記第2のクラスである学習サンプルの数との比に基づいて決定される、サンプルの帰属クラス予測方法。
- [7] 請求項5に記載の方法において、前記第1の個数と前記第2の個数は、前記ステップa)における学習サンプルセット中の全サンプル数と、前記ステップf)における繰り返しの回数の最大値として予め設定した値に基づいて決定される、サンプルの帰属クラス予測方法。
- [8] 請求項1に記載の方法において、前記各学習サンプルは化合物であり、前記第1と第2のクラスは、任意の毒性を有する化合物のクラスと前記毒性を有さない化合物のクラスである、サンプルの帰属クラス予測方法。
- [9] a) 学習サンプルセットの各学習サンプルに対して発生させた複数の説明変数に基づいて、前記各学習サンプルを第1または第2のクラスに分類予測する判別関数を求める手順と、
b) 前記判別関数に基づいて各学習サンプルと未知サンプルの判別得点を算出する手順と、
c) 前記未知サンプルの判別得点が、前記全ての学習サンプルのうちの最大の判別得点以上あるいは最小の判別得点以下の何れかであるか否かを判定する手順と、
d) 前記c)で肯定判断の場合、前記未知サンプルの判別得点に従ってその帰属クラスを決定する手順と、
e) 前記c)で否定判断の場合、少なくとも、前記最大の判別得点を有する学習サンプルと前記最小の判別得点を有する学習サンプルとを前記学習サンプルセットから除去して、残りの学習サンプルによって新たな学習サンプルセットを構築する手順と、
f) 前記新たな学習サンプルセットに対して、前記a)～e)の手順を繰り返して実行する手順と、をコンピュータに実行させるための、サンプルの帰属クラス予測プログラム。

- [10] 請求項9に記載のプログラムにおいて、さらに、g) 帰属クラス未定の未知サンプルが存在するか否かを判定する手順を備え、当該手順で否定判断の場合、前記手順f)の繰り返しを停止する、サンプルの帰属クラス予測プログラム。
- [11] 請求項9に記載のプログラムにおいて、前記手順a)における判別関数は、前記複数の説明変数に特徴抽出を実施して不要な説明変数を除去した残りの説明変数に基づいて求める、サンプルの帰属クラス予測プログラム。
- [12] 請求項9に記載のプログラムにおいて、前記手順f)の繰り返しは、前記学習サンプルセット中の総サンプル数と、前記特徴抽出後の説明変数の数との比が予め決定した一定数以下となった場合、停止される、サンプルの帰属クラス予測プログラム。
- [13] 請求項9に記載のプログラムにおいて、前記手順e)では、最大の判別得点を有する学習サンプルから判別得点の大きい順に第1の個数の学習サンプルを、最小の判別得点を有する学習サンプルから判別得点の小さい順に第2の個数の学習サンプルを、前記学習サンプルセットよりそれぞれ除去する、サンプルの帰属クラス予測プログラム。
- [14] 請求項13に記載のプログラムにおいて、前記第1の個数と前記第2の個数は、前記学習サンプルセットにおいて、実測値が前記第1クラスである学習サンプルの数と前記第2のクラスである学習サンプルの数との比に基づいて決定される、サンプルの帰属クラス予測プログラム。
- [15] 請求項13に記載のプログラムにおいて、前記第1の個数と前記第2の個数は、前記手順a)における学習サンプルセット中の全サンプル数と、前記手順f)における繰り返しの回数の最大値として予め設定した値とに基づいて決定される、サンプルの帰属クラス予測プログラム。
- [16] 請求項9に記載のプログラムにおいて、前記各学習サンプルは化合物であり、前記第1、第2のクラスは、任意の毒性を有する化合物のクラスと前記毒性を有さない化合物のクラスである、サンプルの帰属クラス予測プログラム。
- [17] クラス1あるいはクラス2に帰属することが既知の複数のサンプルで構成される学習サンプルセットをデータ解析して、前記クラス1あるいはクラス2への帰属が不明である未知サンプルの帰属クラスを予測する装置であって、

前記学習サンプルセットの各学習サンプルと前記未知サンプルに対して説明変数を発生する説明変数発生装置と、

前記各学習サンプルについて発生した説明変数に基づいて判別分析を行い、上記クラス1およびクラス2を判別するための判別関数を得る、判別関数作成エンジンと、

前記作成された判別関数に基づいて、前記各学習サンプルと前記未知サンプルの判別得点を算出する、判別得点算出装置と、

前記算出された各学習サンプルの判別得点と前記未知サンプルの判別得点を比較する、判別得点比較装置と、

前記判別得点比較装置の比較結果に基づいて、前記未知サンプルの帰属クラスを決定する帰属クラス決定装置と、

前記判別得点比較装置の比較結果に基づいて、前記学習サンプルセットから、少なくとも最大の判別得点を有する学習サンプルと最小の判別得点を有する学習サンプルとを取り除いて新たな学習サンプルセットを構築するサンプルセット設定装置と、

前記設定された新たな学習サンプルセットを前記学習サンプルセットとして用いて前記説明変数発生装置、前記判別関数作成エンジン、前記判別得点算出装置、前記判別得点比較装置、前記帰属クラス決定装置および前記サンプルセット設定装置を繰り返し動作させる制御装置と、を備え、

前記帰属クラス決定装置は、前記判別得点比較装置における比較結果が、前記未知サンプルの判別得点が前記各学習サンプルのうちの最大の判別得点以上か、あるいは最小の判別得点以下かの何れかである場合、前記未知サンプルの判別得点にしたがって前記未知サンプルの帰属クラスを決定する、サンプルの帰属クラス予測装置。

[18] 請求項17に記載の装置において、更に、分類予測に不要な説明変数を除去するために、前記説明変数発生装置によって発生された説明変数に対して特徴抽出を行うための特徴抽出エンジンを備える、サンプルの帰属クラス予測装置。

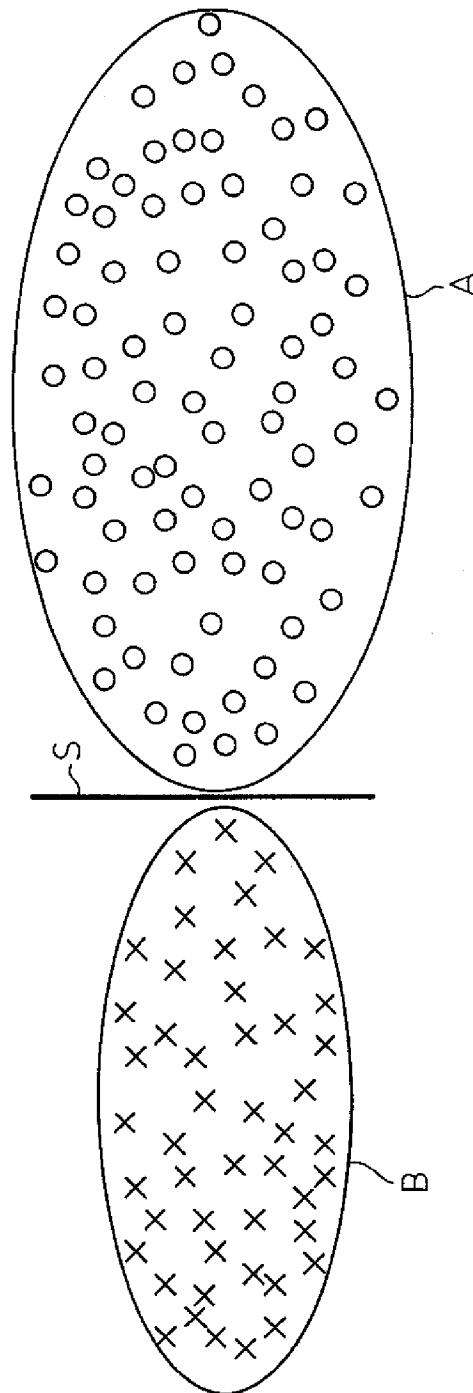
[19] 請求項17に記載の装置において、さらに、前記各学習サンプルおよび未知サンプルとして化合物の構造式を入力する入力装置を備え、前記説明変数発生装置は、

入力された構造式に基づいて複数の構造記述子を前記説明変数として発生させる、
サンプルの帰属クラス予測装置。

- [20] 請求項19に記載の装置において、さらに、前記各学習サンプルとして入力された
構造式と前記各学習サンプルの既知の帰属クラスとをデータとして保存するサンプル
ライブラリを備える、サンプルの帰属クラス予測装置。

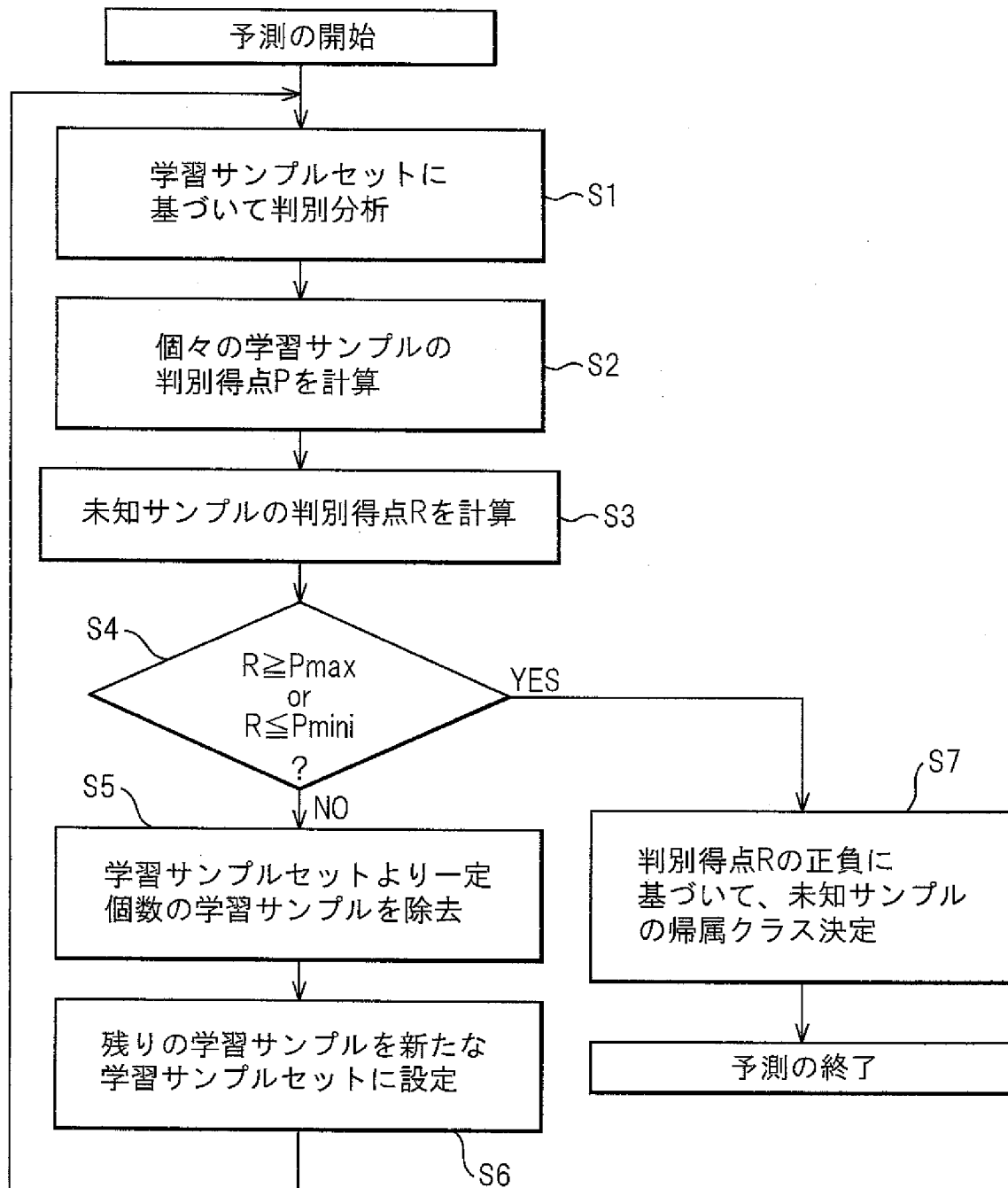
[図1]

Fig.1



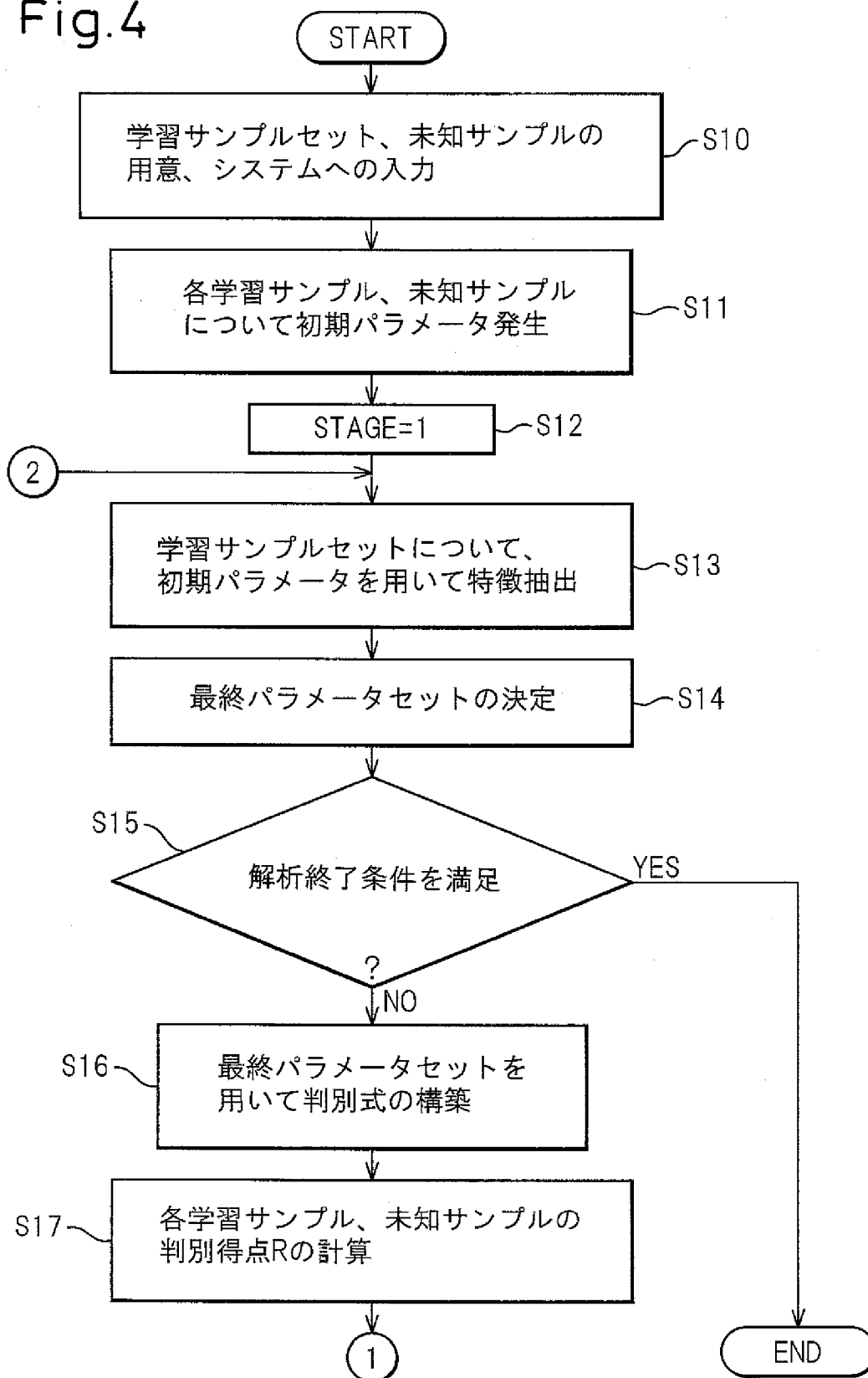
[図3]

Fig.3

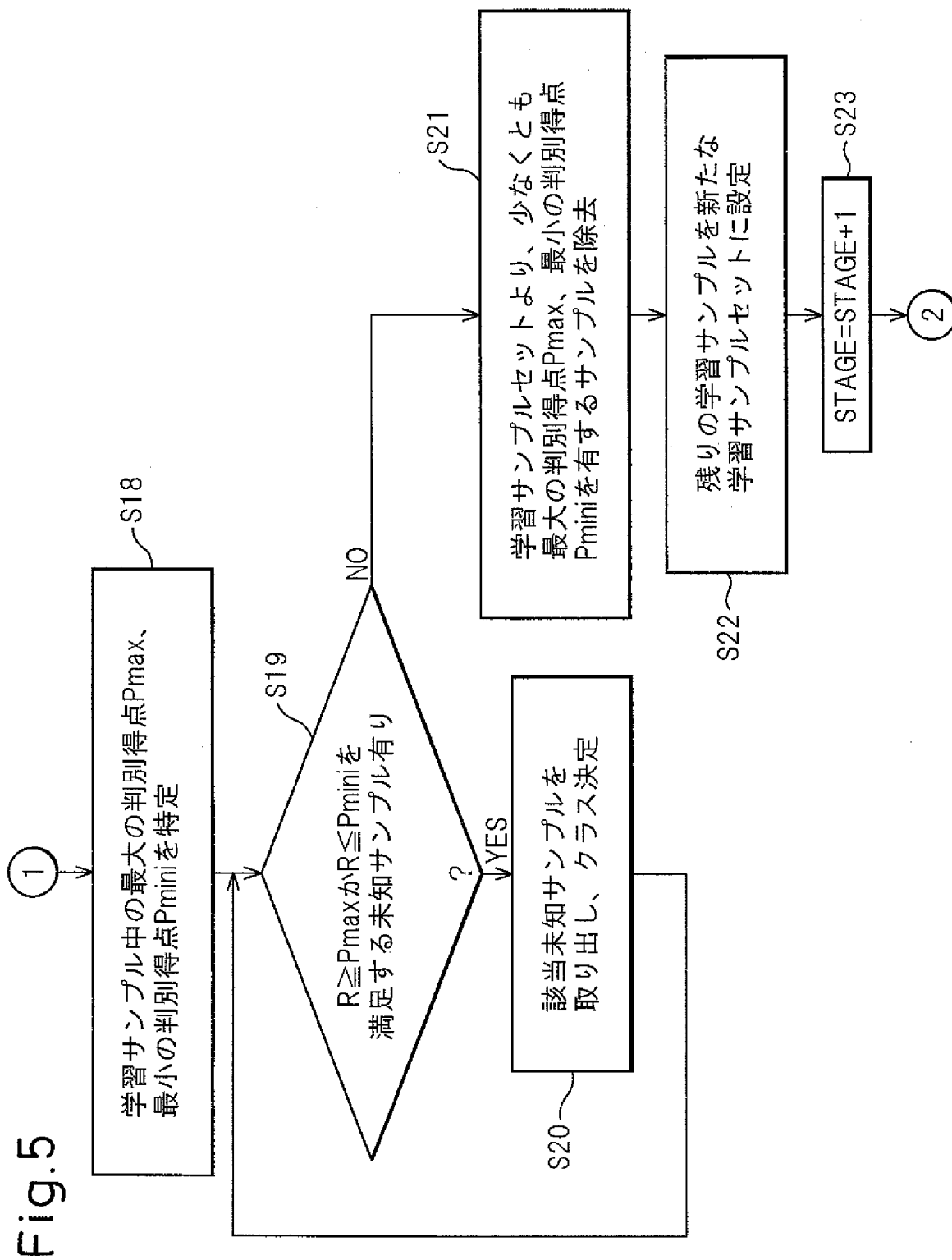


[図4]

Fig.4

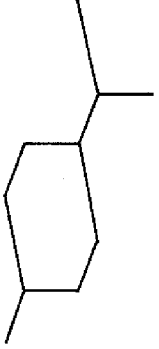
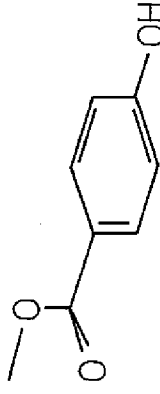
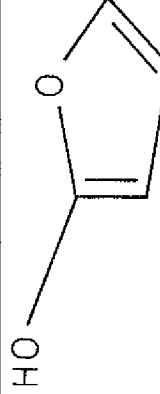
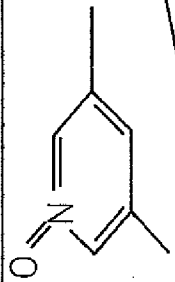


[図5]



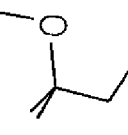
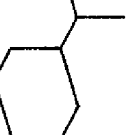
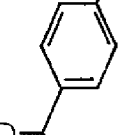
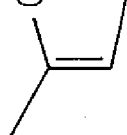
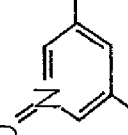
[図6]

Fig. 6

63	A Structure	B CAS Number	C Ames test
1		994-05-8	nonmutagen (-)
2		99-82-1	nonmutagen (-)
3		99-76-3	nonmutagen (-)
4		98-00-0	mutagen (+)
5		97-56-1	mutagen (+)

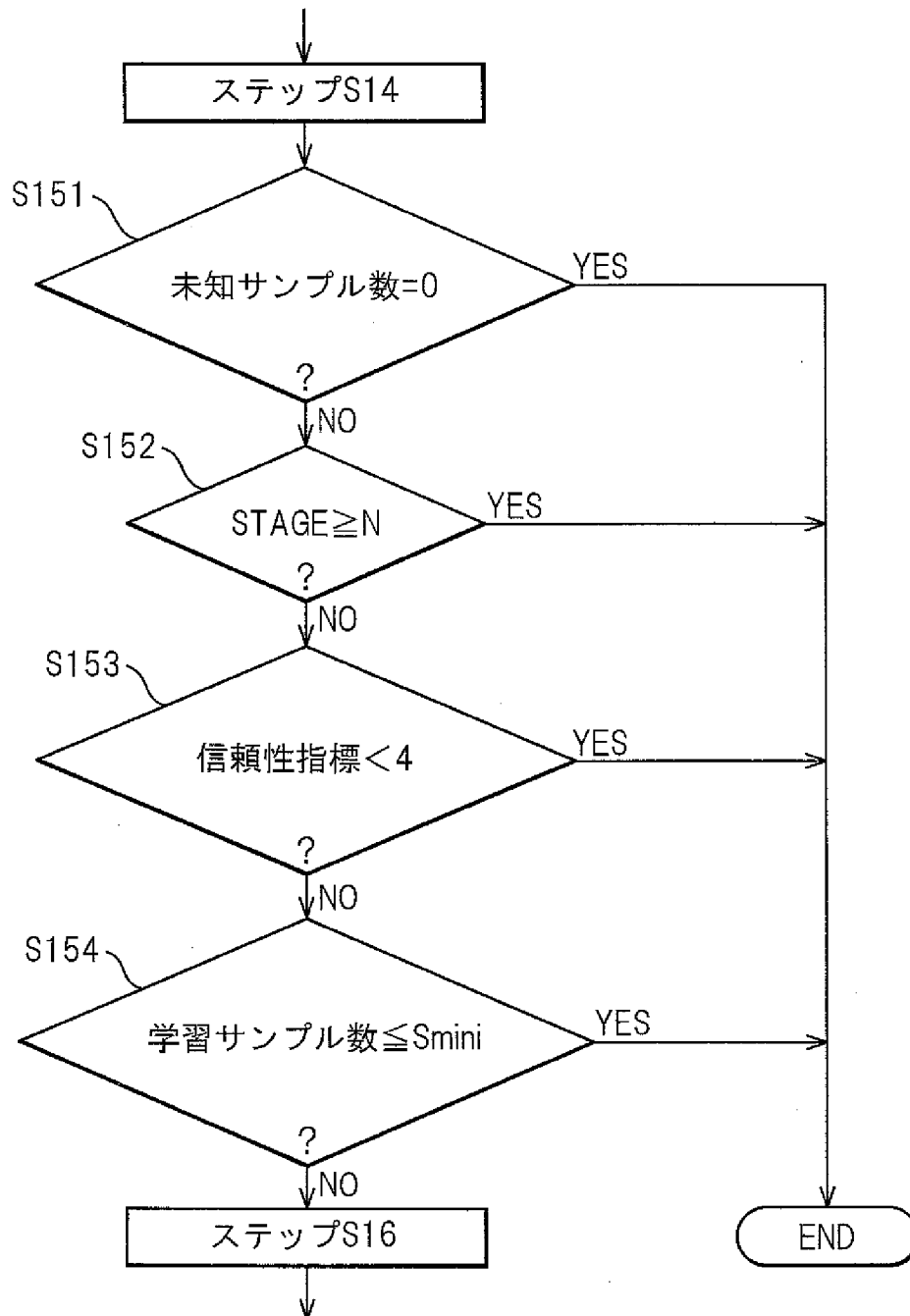
[図7]

Fig. 7

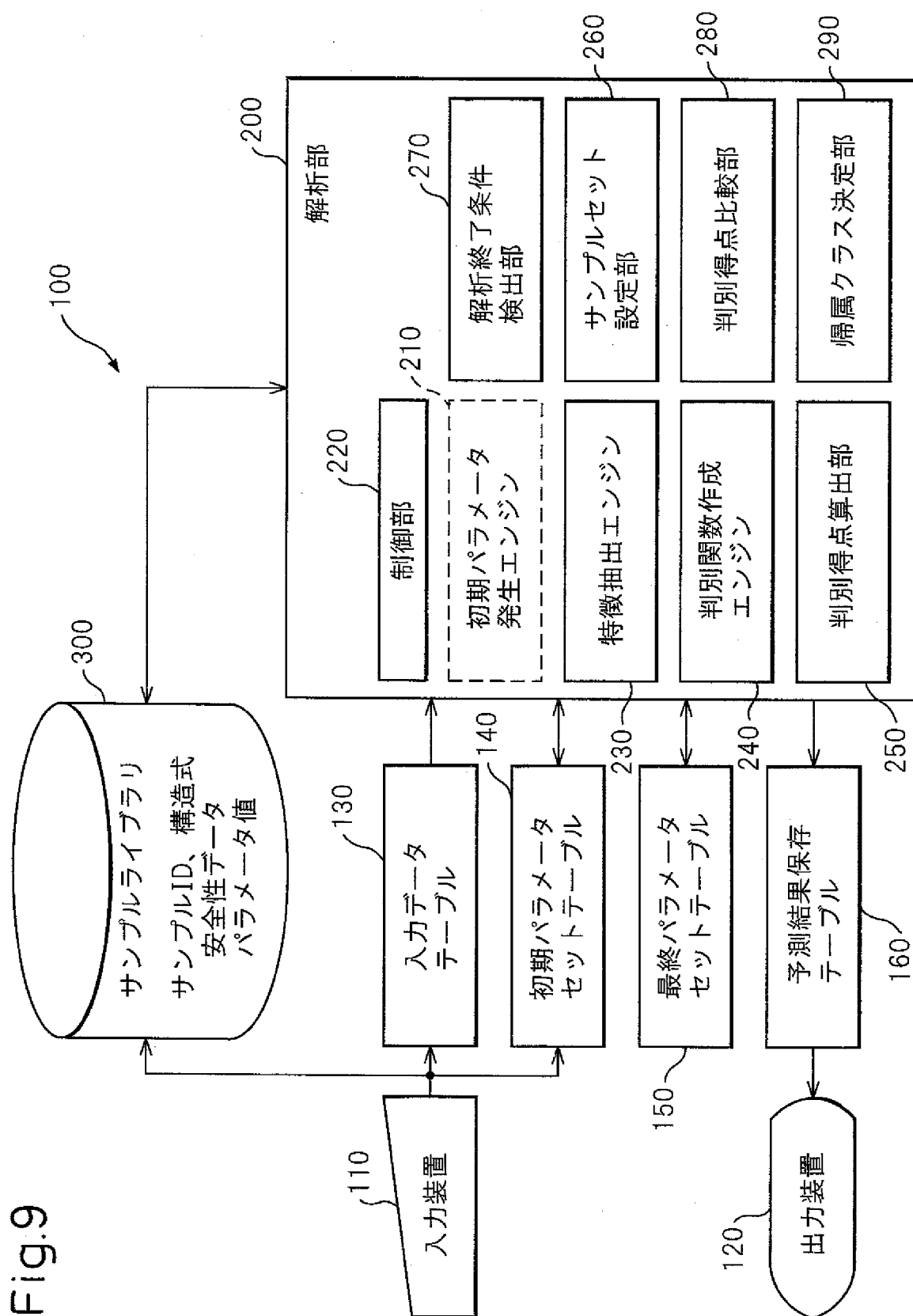
77	70	71	72	73	74	75	76	77
	A Structure	B Molecular Mass (Whole Molecule)	C Molecular Surface Area (Whole Molecule)	D Molecular Volume (Whole Molecule)	E Kappa 1 index (Whole Molecule)	F log P (Whole Molecule)	G Shape Flexibility index (Whole Molecule)	H Randic Topological index (Whole Molecule)
1		102.2 (x11)	143.704 (x21)	91.8144 (x31)	7	1.3154	2.30088	3.12132
2		140.3 (x12)	190.412 (x22)	129.174 (x32)	8.1	3.7656	2.76071	4.69838
3		152.16 (x13)	172.868	101.168	9.09091	1.4923	2.35543	5.23638
4		98.11 (x14)	120.712	68.9598	5.14286	0.4597	1.24264	3.43185
5		123.17 (x15)	152.166	89.6854	7.11111	0.9371	1.48936	4.18154

[図8]

Fig.8



[図9]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2008/064058

A. CLASSIFICATION OF SUBJECT MATTER

G06N5/04(2006.01)i, C40B30/02(2006.01)i, G06Q10/00(2006.01)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)

G06N5/04, C40B30/02, G06Q10/00

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 1996-2008

Kokai Jitsuyo Shinan Koho 1971-2008 Toroku Jitsuyo Shinan Koho 1994-2008

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.
X	JP 2002-73076 A (Sumitomo Metal Industries, Ltd.), 12 March, 2002 (12.03.02), Par. Nos. [0087] to [0136] (Family: none)	1-3, 5-7, 9-11, 13-15, 17, 18
Y		4, 8, 12, 16, 19, 20
Y	JP 2004-86897 A (Fuji Electric Holdings Co., Ltd.), 18 March, 2004 (18.03.04), Par. Nos. [0044] to [0052] (Family: none)	4, 12
Y	JP 8-123781 A (Fujitsu Ltd.), 17 May, 1996 (17.05.96), Par. Nos. [0009] to [0011] & US 5796632 A1	4, 8, 12, 16, 19, 20

☒ 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
04 November, 2008 (04.11.08)Date of mailing of the international search report
18 November, 2008 (18.11.08)Name and mailing address of the ISA/
Japanese Patent Office

Authorized officer

Facsimile No.

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2008/064058

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 5-143636 A (Omron Corp.) , 11 June, 1993 (11.06.93) , Par. Nos. [0027] to [0039] (Family: none)	1-20

A. 発明の属する分野の分類（国際特許分類（I P C））

Int.Cl. G06N5/04(2006.01)i, C40B30/02(2006.01)i, G06Q10/00(2006.01)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（I P C））

Int.Cl. G06N5/04, C40B30/02, G06Q10/00

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	1 9 2 2 - 1 9 9 6 年
日本国公開実用新案公報	1 9 7 1 - 2 0 0 8 年
日本国実用新案登録公報	1 9 9 6 - 2 0 0 8 年
日本国登録実用新案公報	1 9 9 4 - 2 0 0 8 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 請求の範囲の番号
X	JP 2002-73076 A（住友金属工業株式会社）2002.03.12, 段落【0087】 －【0136】（ファミリーなし）	1-3, 5-7, 9-11 , 13-15, 17, 18
Y		4, 8, 12, 16, 19 , 20
Y	JP 2004-86897 A（富士電機ホールディングス株式会社） 2004.03.18, 段落【0044】－【0052】（ファミリーなし）	4, 12
Y	JP 8-123781 A（富士通株式会社） 1996.05.17, 段落【0009】－【0011】 & US 5796632 A1	4, 8, 12, 16, 19 , 20

☒ 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

0 4 . 1 1 . 2 0 0 8

国際調査報告の発送日

1 8 . 1 1 . 2 0 0 8

国際調査機関の名称及びあて先

日本国特許庁（I S A / J P）

郵便番号 1 0 0 - 8 9 1 5

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）

北川 純次

5 B

3 6 5 0

電話番号 0 3 - 3 5 8 1 - 1 1 0 1 内線 3 5 4 5

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 請求の範囲の番号
A	JP 5-143636 A (オムロン株式会社) 1993.06.11, 段落【0027】－【0039】 (ファミリーなし)	1-20