(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
2009 年4 月2 日 (02.04.2009)

PCT

(10) 国際公開番号
WO 2009/041101 A1

(51) 国際特許分類:
G06F 17/30 (2006.01) *G06N 3/00* (2006.01)

(21) 国際出願番号: PCT/JP2008/057705

(22) 国際出願日: 2008 年4 月21 日 (21.04.2008)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願2007-253703 2007 年9 月28 日 (28.09.2007) JP

(71) 出願人 (米国を除く全ての指定国について): 日本電気株式会社 (NEC CORPORATION) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 Tokyo (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 藤巻 遼平 (FUJIMAKI, Ryohei) [JP/JP]; 〒1088001 東京都港区芝五丁目７番１号 日本電気株式会社内 Tokyo (JP).

(74) 代理人: 工藤 実 (KUDOH, Minoru); 〒1400013 東京都品川区南大井六丁目２４番１０号カドヤビル６階 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, 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,

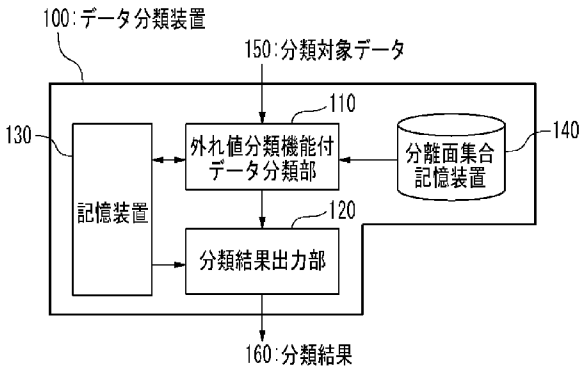
[続葉有]

WO 2009/041101 A1

(54) Title: METHOD FOR CLASSIFYING DATA AND DEVICE FOR CLASSIFYING DATA

(54) 発明の名称: データ分類方法およびデータ分類装置

[図1]



100 DATA CLASSIFICATION DEVICE
150 CLASSIFICATION OBJECT DATA
130 STORAGE DEVICE
110 DATA CLASSIFICATION SECTION WITH OUTLIER CLASSIFICATION FUNCTION
120 CLASSIFICATION RESULT OUTPUT SECTION
160 CLASSIFICATION RESULT
140 SEPARATION FACE SET STORAGE DEVICE

(57) Abstract: A separation face set storage section stores information for defining a plurality of separation faces each separating a feature space into at least one known class region and unknown class region corresponding to at least one known class, respectively. Respective known class regions are separated by two or more nonintersecting separation faces. A data classification device determines classification of classification object data by performing calculation for determining to which of the at least one known class region and unknown class region the classification object data capable of calculating the inner product in the feature space belongs. The data classification method and a data classification device for performing highly reliable identification and outlier classification simultaneously by the same procedure are thereby provided.

(57) 要約: 分離面集合記憶部は、特徴空間を少なくとも１つの既知クラスにそれぞれ対応する少なくとも１つの既知クラス領域と未知クラス領域とに分離する複数の分離面を規定する情報を記憶する。各既知クラス領域は、互いに交差しない２以上の分離面によって分離される。データ分類装置は、特徴空間における内積が計算可能である分類対象データが、少なくとも１つの既知クラス領域と未知クラス領域とのうちのどの領域に属するかを計算することによって、分類対象データの分類を決定する。信頼性の高い識別と外れ値分類を同じ手順で同時に行うことができるデータ分類方法およびデータ分類装置が提供される。

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, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:
— 国際調査報告書

明 細 書

データ分類方法およびデータ分類装置

技術分野

[0001] 本発明は、データ分類方法およびデータ分類装置に関し、特に、複数の分離面を利用することによって既知のクラスおよび外れ値を同時に分類可能なデータ分類方法およびデータ分類装置に関する。この出願は、2007年9月28日に出願された日本特許出願2007－253703号を基礎とする。その日本特許出願の開示はこの参照により、ここに取り込まれる。

背景技術

[0002] データ分類は、未分類データが与えられた場合に、該データの属するクラスを推定する技術であり、データ解析の最も基本的な要素の一つである。特に、クラス間の分離面など、特徴空間を複数の領域に分ける分離面を利用したデータ分類技術は、モデル表現力が高い。そのため、画像データ、たんぱく質や遺伝子データをはじめとしたデータ分類はもちろん、クラスラベルを故障情報とした場合には故障診断、インターネットやソーシャルネットワークなどネットワーク間のリンクの有無をクラスラベルとした場合にはリンクの予測など幅広い問題およびデータ構造に対して応用が可能である。

[0003] 分離面を利用したデータ分類方法は、識別と外れ値分類の2つの技術に大別できる。前者は、クラスラベルの付いたデータからクラスを分離する分離面を学習し、分類対象データを既知のクラスへ分類する技術である。後者は学習データを1つのクラスとみなし、学習データの分布する領域とそれ以外の領域を分離する分離面を学習することで、分類対象データが該クラスに属するか、該クラスからは外れるかを分類する技術である。また、識別および外れ値分類を同時に実施するデータ分類方法としては、分離面を利用したデータ分類方法の組み合わせとして容易に類推することができる方法が幾つかある。

[0004] まず、学習データに関するクラスの数が1つの場合、データ分類は外れ値分類となるため、1クラスサポートベクトルマシン（文献5第8章、文献3）などの公知の外れ値分

類技術を利用することが考えられる。

[0005] 次に、学習データに関するクラスの数が2つ以上の場合、1クラスサポートベクトルマシンなどの外れ値分類方法を、各クラスに対して個々に学習し、分類対象データに対して全クラスに対して外れ値であると判定された場合には外れ値とし、1つまたは複数のクラスに対してそのクラスに属すると判定された場合には、それらのクラスの1つまたは複数に分類する方法が考えられる。

[0006] 学習データに関するクラスの数が2つ以上の場合の他の方法として、1クラスサポートベクトルマシンなどの外れ値分類方法と、サポートベクトルマシン(文献1、文献2、文献6)などの分離面を利用した識別方法を組み合わせ、まず全クラスをまとめて外れ値分類方法によって学習し、次に既知のクラスに関する識別方法を学習する方法が考えられる。この方法では、まず分類対象データを外れ値検出方法によって外れ値かどうかを判定し、外れ値ではない場合に識別方法によって、既知のどのクラスに属するかを分類する。

[0007] 他方、複数の分離面を利用した技術としては、多クラスサポートベクトルマシンがある。多クラスサポートベクトルマシンの実現方法は幾つかあるが、2クラスのサポートベクトルマシンをクラスの組み合わせごとに計算し多数決をとる方法と、文献7および文献4で提案する方法のように複数の超空間を同時に最適化する方法がある。

[0008] 以下に文献のリストを挙げる。

[文献1]特開2007－115245号公報

[文献2]特開2007－95069号公報

[文献3]特開2005－345154号公報

[文献4]特開2007－52507号公報

[文献5]Bernhard Scholkopf and Alex Smola. Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond. MIT Press. 2002.

[文献6]Bernhard Scholkopf, Alex J. Smola, Robert C. Williamson and Peter L.Bartlett. New Support Vector Algorithms. Neural Computation. Vol.12: page1207-1245. 2000.

[文献7]Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, Yasemin Alt

un. Large Margin Methods for Structured and Interdependent Output Variables. Journal of Machine Learning Research Vol.6: page 1453-1484. 2005.

[文献8]A.L. Yuille and A. Rangarajan. The concave-convex procedure. Neural Computation. Vol.15: page 915-936. 2003.

発明の開示

[0009] 従来の識別および外れ値分類を同時に実施するデータ分類方法には以下のような課題がある。

[0010] まず、1クラスサポートベクトルマシンやサポートベクトルマシンのような単一の分離面によってデータを分類する場合、データの片側の境界面のみを考慮することになり、逆側の境界を考慮できないため、分類が楽観的になるという問題点がある。

[0011] その理由は、図18に示されるように、分離超平面(単に超平面とも称す)を利用した1クラスサポートベクトルマシンでは、データの片面の分離境界のみを考慮し、逆側の境界は考慮されないためである。また、図19に示されるように、分離超球面(単に超球面とも称す)を利用した1クラスサポートベクトルマシンでは、データの外側の分離境界のみを考慮し、内側の境界は考慮されないためである。これはその他の、分離面を利用した公知のデータ分類装置に共通する課題である。

[0012] さらに、分離面を利用した公知のデータ分類技術を組み合わせた場合には、データ分類精度の信頼性が低下するという問題点がある。

[0013] その理由は、各クラスに対する外れ値分類を組み合わせた場合には、各クラスを独立に扱うため、クラス間の関係を考慮しないためである。また、外れ値分類と識別を組み合わせる場合には、異なるクラスを1つのクラスと考えるため、外れ値分類の精度が低下するためである。これは、上記の組み合わせ方法以外の組み合わせ方をした場合にも起こりうる課題である。

[0014] これら公知の技術を組み合わせた場合には、複数の分離面を利用しているが、それらは独立に計算され利用されているため、1つずつ分離面を利用していることと同義である。

[0015] さらに、従来の分離面を利用したデータ分類方法には、外れ値分類と識別を同時に行うという考え方がないため、同じモジュールによって外れ値分類と識別を同時に

行うことができないという課題もある。

[0016] さらに、多クラスサポートベクトルマシンは複数の分離面を利用するが、外れ値分類を行うことができないという課題がある。

[0017] その理由は、多クラスサポートベクトルマシンでは、既知クラス間を分類する分離面のみを考慮し、未知のクラスと既知クラスの境界を考慮しないためである。換言すれば、既知のクラスは1つの分離面を挟んで他の既知のクラスと接しており、既知のクラスの間に未知のクラスを介在させるという考えがない。

[0018] 本発明の目的は、信頼性の高い識別と外れ値分類を同じ手順で同時に行うことができるデータ分類方法およびデータ分類装置を提供することである。

[0019] 本発明の一実施形態におけるデータ分類装置は、特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を規定する情報を記憶する分離面集合記憶部を備える。少なくとも1つの既知クラス領域の各々は複数の分離面のうちの互いに交差しない2以上によって外部領域と分離される。データ分類装置は更に、内積が計算可能な分類対象データが、分離面記憶部に記憶された情報で規定される少なくとも1つの既知クラス領域と未知クラス領域とのうちのどの領域に属するかを計算することによって、分類対象データの分類を決定する分類部を備える。

[0020] 本発明の一実施形態におけるデータ分類方法は、(a)特徴空間における内積が計算可能な分類対象データを入力する工程と、(b)特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を分離面記憶部から入力する工程とを備える。少なくとも1つの複数の既知クラス領域の各々は複数の分離面のうちの互いに交差しない2以上によって外部領域と分離される。データ分類方法は更に、(c)分類対象データが、少なくとも1つの既知クラス領域と未知クラス領域とのうちのどの領域に属するかを計算することによって、分類対象データの分類を決定する工程を備える。

[0021] 本発明の一実施形態における分離面集合計算装置は、特徴空間における内積が計算可能であり少なくとも1つの既知クラスのいずれかにそれぞれ分類されている複数の学習データを記憶する学習データ記憶部と、学習データ記憶部に記憶された複

数の学習データおよび複数の学習データの各々の分類に基づいて、特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を計算する分離面集合計算部とを備える。少なくとも1つの既知クラス領域の各々は複数の分離面のうちの互いに交差しない2以上によって外部領域と分離される。分離面集合計算装置は更に、複数の分離面を規定する情報を記憶する分離面集合記憶部を備える。

[0022] 本発明の一実施形態におけるプログラムは、以下の(a)～(c)を備える方法をコンピュータに実行させる。

(a)特徴空間における内積が計算可能な分類対象データを入力する工程。

(b)特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を分離面記憶部から入力する工程。少なくとも1つの複数の既知クラス領域の各々は複数の分離面のうちの互いに交差しない2以上によって外部領域と分離される。

(c)分類対象データが、少なくとも1つの既知クラス領域と未知クラス領域とのうちのどの領域に属するかを計算することによって、分類対象データの分類を決定する工程。

[0023] 本発明の一実施形態におけるプログラムは、以下の(a)～(c)を備える方法をコンピュータに実行させる。

(a)特徴空間における内積が計算可能であり少なくとも1つの既知クラスのいずれかにそれぞれ分類されている複数の学習データを記憶する工程。

(b)学習データ記憶部に記憶された複数の学習データおよび複数の学習データの各々の分類に基づいて、特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を計算する工程。少なくとも1つの既知クラス領域の各々は複数の分離面のうちの互いに交差しない2以上によって外部領域と分離される。

(c)複数の分離面を規定する情報を記憶する工程。

[0024] 本発明によれば信頼性の高い識別と外れ値分類を同じ手順で同時に行うことができる。識別と外れ値分類を同じ手順で同時に行うことができる理由は、内積が計算可能であり1以上の既知クラスに分類されている特徴空間における複数の学習データ

および複数の学習データの分類に基づいて、特徴空間を1以上の既知クラスにそれぞれ対応する1以上の既知クラス領域と未知クラス領域とに分離する複数の分離面であって、1クラス当たり2個以上で且つ互いに交差しない複数の分離面を計算し、分類が未知である前記特徴空間における内積が計算可能な分類対象データの分類時には、分類対象データが、複数の分離面によって1以上のクラス領域とそれ以外の未知クラス領域とに分離された特徴空間内のうちのどの領域に属するかを計算することによって、その分類対象データの分類を決定するためである。また、信頼性の高いデータ分類を行える理由は、それぞれの既知クラスは2以上の分離面によって境界が定められているため、1つの分離面によって境界を定める場合と比較して、データ分類の信頼性がより高まるからである。

図面の簡単な説明

[0025] [図1]本発明の第1の実施の形態に係るデータ分類装置の構成を示すブロック図である。

[図2]本発明の第1の実施の形態に係る超平面を利用したデータ分類の一例である。

[図3]本発明の第1の実施の形態に係る超球面を利用したデータ分類の一例である。

[図4]本発明の第1の実施の形態に係る超平面を規定するデータの記憶方法の一例である。

[図5]本発明の第1の実施の形態に係る超球面を規定するデータの記憶方法の一例である。

[図6]本発明の第1の実施の形態に係るデータ分類装置の処理例を示すフローチャートである。

[図7]本発明の第2の実施の形態に係るデータ分類装置の構成を示すブロック図である。

[図8]本発明の第2の実施の形態に係る分離面集合計算装置の構成を示すブロック図である。

[図9]本発明の第3の実施の形態に係るデータ分類装置の構成を示すブロック図である。

[図10]本発明の第3の実施の形態に係る超平面集合計算装置の構成を示すブロック

図である。

[図11]本発明の第3の実施の形態に係るデータ分類装置によって、クラスの数が1つの場合に計算されるデータ分類の概念図である。

[図12]本発明の第3の実施の形態に係るデータ分類装置によって、クラスの数が2つの場合に計算されるデータ分類の概念図である。

[図13]本発明の第3の実施の形態に係るデータ分類装置によって、クラスの数が3つ以上の場合に計算されるデータ分類の概念図である。

[図14]本発明の第3の実施の形態に係るデータ分類装置で使用するのが好ましくない超平面の説明図である。

[図15]本発明の第4の実施の形態に係るデータ分類装置の構成を示すブロック図である。

[図16]本発明の第4の実施の形態に係る超球面集合計算装置の構成を示すブロック図である。

[図17]本発明の第4の実施の形態に係るデータ分類装置によって計算されるデータ分類の概念図である。

[図18]本発明に関連する、超平面を利用したデータ分類技術の例である。

[図19]本発明に関連する、超球面を利用したデータ分類技術の例である。

発明を実施するための最良の形態

[0026] 次に、本発明の実施の形態について、図面を参照して詳細に説明する。

[0027] [第1の実施の形態]

図1を参照すると、本発明の第1の実施の形態に関わるデータ分類装置100は、外れ値分類機能付データ分類部110と、分類結果出力部120と、記憶装置130と、分離面集合記憶装置140とを備えている。データ分類装置100は、パーソナルコンピュータなどのコンピュータによって実現することができる。その場合、外れ値分類機能付データ分類部110と分類結果出力部120とは、記憶装置に格納されたプログラムをCPUなどの処理装置が読み出してそのプログラムに既述された手順に従って実行動作を行うことにより実現される。

[0028] このデータ分類装置100は、分類対象データ150を入力し、分類対象データ150

が複数の分離面によって1以上のクラス領域（既知クラス領域）とそれ以外の未知クラス領域とに分離された特徴空間内のどの領域に属するかを計算することによって、分類対象データ150を既知のどのクラスに分類すべきか、または外れ値に分類すべきかを推定し、その推定結果を分類結果160として出力する。

[0029] 分類対象データ150は、分類が未知のベクトルデータである。今、分類対象データ150に含まれる属性の数をdとし、分類対象データ150を式（1）のようにd次元のベクトルとして表現する。式（1）において、右辺の括弧の右肩に付した記号´は転置を示す（記号´の代わりに記号^{T}を使う場合もある）。また、x^{j}は分類対象データ150のj番目の属性を表し、実数値であっても良いしシンボル値であってもよい。また、xから特徴空間への写像をφとし、xの特徴空間における像をφ（x）と表す。以下で、分類対象データといった場合には、分類対象データと特徴空間における像のどちらを指してもかまわないものとする。

[0030]

$$x = \left(x^{1}, \ldots, x^{j}, \ldots, x^{d}\right)' \qquad (1)$$

[0031] 分離面集合記憶装置140は、特徴空間を1以上の既知クラスにそれぞれ対応する1以上のクラス領域とそれ以外の未知クラス領域とに分離する複数の分離面を規定する情報を記憶する。分離面は、図2に示される超平面A～Dのように特徴空間上で平面を成すものであっても良いし、図3に示される超球面E～Hのように特徴空間上で球面を成すものであっても良く、その他、超円柱面、超錐面などであっても良い。ただし、図2に示される互いに平行な超平面A～D、図3に示される同心の超球面E～Hのように、複数の分離面は互いに交差しないことが必要である。また、図2ではクラス1の領域が2つの超平面A，Bにより、クラス2の領域が2つの超平面C，Dにより、図3におけるクラス3の領域が2つの超球面E，Fにより、クラス4の領域が2つの超球面G，Hにより、それぞれ境界が定められている。このように、既知の1クラス当たり2個以上の分離面によって各既知クラスの境界が定められている。

[0032] 分離面集合記憶装置140に記憶する情報は、分離面を特定する情報であればどのような情報であっても良い。例えば、特徴空間のi番目の基底関数をφiとすると、特徴空間における分離面は基底関数を利用して記述することが可能である。例えば

、分離面が$\Sigma w_i \phi i \phi(x) + b = 0$で表される超平面の場合には、基底$\phi i$および基底の重み$w_i$、切片bを、超平面を規定する情報として記憶すればよい。この際、基底ϕiは全ての超平面で共通のため、例えば図4のように、重みw_iと切片bを表形式で各超平面毎に記憶し、基底ϕiは共通に記憶しておくことが可能である。また、超球面の場合には、中心をc、半径をrとすると、$|\phi(x) - c|^2 = r$と表され、また中心cは特徴空間内の点のため、$c = \Sigma w_i \phi i$と表される。そのため、重みw_iと半径rを図5のように表形式で各超球面毎に保存し、基底ϕiは共通に記憶しておくことが可能である。また、基底関数に関しては、任意の基底関数を利用することが可能であるが、頻繁に利用される基底としては、例えばxの元空間における基底やカーネル関数などが挙げられる。その場合、基底同士の内積が定義されているものとする(カーネル関数とは、特定の条件を満たす任意の基底関数に関する内積を与える関数)。

[0033] 記憶装置130には、分類対象データ150と分離面集合記憶装置140に記憶された複数の分離面との位置関係から、分類対象データ150を分類するためのルールが記憶されている。例えば図2に示されるように複数の超平面によってデータを分類する場合、記憶装置130には、例えば「超平面Aより負の方向→外れ値へ分類」、「超平面Cより正の方向かつ超平面Dより負の方向→クラス2へ分類」などといったルールが記憶されている。また、図3に示されるように複数の超球面によってデータを分類する場合、記憶装置130には、例えば「超球面Eより内側→外れ値へ分類」、「超球面Gより外側かつ超球面Hより内側→クラス4へ分類」などといったルールが記憶されている。この例では超平面と超球面の場合を説明したが、前述したように分離面はその二つに限定されるものではない。分離面としてその他の形状の超曲面を利用することも可能であるし、また異なる種類の分離面を組み合わせることも可能である。なお、記憶装置130には外れ値分類機能付データ分類部110によって判定された分類結果を記憶しておくことも可能である。

[0034] 外れ値分類機能付データ分類部110は、分類対象データ150および分離面集合記憶装置140に記憶された複数の分離面に関する情報を読み込み、分類対象データ150と複数の分離面との位置関係を計算する。分離面は前述したように、例えば超平面、超球面、超円柱面、超錐面などである。位置関係とは、例えば超平面の場合

にはデータが超平面上、正の側、負の側のどの位置にあるかであり、超球面の場合には超球面上、超球面の内側、超球面の外側のどの位置にあるかのことである。この位置関係からデータを分類するルールが前述したように記憶装置130に保存されており、外れ値分類機能付データ分類部110は、位置関係および分類ルールを利用してデータを分類する。

[0035] 分類結果出力部120は、外れ値分類機能付データ分類部110で判定された分類結果を、外れ値分類機能付データ分類部110から直接受け取るか、記憶装置130に記憶された分類結果を読み出して出力する。出力先は、データ分類装置100に接続されたディスプレイ等の出力装置であっても良いし、ネットワークを介して接続された出力装置や端末装置であっても良い。

[0036] 次に本実施の形態に関わるデータ分類装置の全体の動作を説明する。

[0037] 図6を参照すると、データ分類装置100の外れ値分類機能付データ分類部110は、d個の属性を含む分類対象データ150を入力し(S100)、また分離面集合記憶装置150から複数の分離面の情報を入力する(S101)。

[0038] 次に、外れ値分類機能付データ分類部110は、入力された分類対象データ150および複数の分離面の情報を利用して、分類対象データ150と複数の分離面との位置関係を計算する(S102)。計算は、例えば図2および図4の超平面Aを例にすると、データxに対して、$\Sigma w_i^A \phi i \phi(x) + b^A$を計算し、その値の位置関係(0、正、負によって、それぞれ超平面A上、超平面Aより正側、超平面Aより負側のいずれかに分類される)を判定することができる。また図3および図5の超球面Eの場合でも、位置関係(データxに対して$|\phi(x) - \Sigma w_i^E \phi i|^2$が$r^E$と等しいか、大きいか、小さいかによって、それぞれ超球面E上、超球面Eより外側、超球面Eより内側のいずれかに分類される)を判定することが可能である。

[0039] 次に、外れ値分類機能付データ分類部110は、記憶装置130に記憶されている分類ルールを読み込み、分類対象データ150がどのクラスに属するかを判定する(S103)。そして、分類結果出力部120は、外れ値分類機能付データ分類部110の分類結果を出力する(S104)。

[0040] データ分類に関しては、既知のクラス数は1つ乃至は複数であり、1つの場合は外

れ値分類を行うデータ分類装置として機能する。

[0041] 次に本実施の形態の効果を説明する。

[0042] 本実施の形態によれば、識別と外れ値分類を同じ手順で同時に行うことができる。その理由は、特徴空間を1以上の既知クラスにそれぞれ対応する1以上のクラス領域とそれ以外の未知クラス領域とに分離する複数の分離面と分類対象データ150との位置関係を計算し、分類対象データ150が1以上のクラス領域とそれ以外の未知クラス領域とのうちのどの領域に属するかを計算することによって、分類対象データ150の分類を決定するためである。

[0043] また本実施の形態によれば、信頼性の高いデータ分類を行うことができる。その理由は、それぞれの既知クラスは2以上の分離面によって境界が定められているため、1つの分離面によって境界を定める場合と比較して、データ分類の信頼性がより高まるからである。

[0044] [第2の実施の形態]

図7を参照すると、本発明の第2の実施の形態に関わるデータ分類装置200は、図1に示した第1の実施の形態に関わるデータ分類装置100と比較して、分離面集合記憶装置140に代えて分離面集合記憶装置210を有する点、および分離面集合計算装置220が接続されている点で相違する。

[0045] 分離面集合計算装置220は、1以上の既知クラスに分類されている複数の学習データおよびその分類に基づいて複数の分離面を計算する。複数の分離面は、特徴空間を1以上の既知クラスにそれぞれ対応する1以上のクラス領域とそれ以外の未知クラス領域とに分離する。1以上のクラス領域の各々は、この複数の分離面のうちの互いに交差しない2以上によって、他の領域と分離される。また、分離面集合記憶装置210は、分離面集合計算装置220で計算された複数の分離面を規定する情報を記憶する装置である。

[0046] 分離面集合計算装置220は、図8に示されるように、分離面集合最適化部221と記憶装置222と分離面集合出力部223とを備える。分離面集合最適化部221は、学習データ記憶装置224から学習用のデータを入力する。分離面集合出力部223は、最適化された分離面集合225を出力する。

[0047] 学習データ記憶装置224には、分類対象データ150と同じ属性を有するデータx_iと、データx_iの属するクラスラベルy_iの組の集合が記憶されている。ここで、iは学習データのインデックスとし、Nを所定の整数とし、学習データはi＝1，…，Nまで入力されるとする。

[0048] 分離面集合最適化部221は、学習データに対する分類誤差の最小化、分離面集合の複雑性の最小化、および各クラス領域の大きさの最小化を同時に最適化する複数の分離面を計算する。利用する複数の分離面に関しては、事前に候補となる分離面の組み合わせを記憶装置222へ記憶しておき、最適化時に記憶装置222からそれらの候補を読み込んで利用しても良い。または任意の分離面の組み合わせに対して最適化を行うことにより最適な分離面集合を選択してもよい。

[0049] 分類誤差は、任意の誤差を利用することが可能であり、例としては誤分類データ数、誤分類データに対する2乗損失、誤分類データに対する絶対値損失、誤分類データに対するヒンジ損失などが挙げられる。

[0050] 分離面集合の複雑性は、任意の複雑性の基準を利用することが可能である。例としては、j番目の分離面をf_jとすると、f_jのL1複雑性$|f_j|$、L2複雑性$|f_j|^2$、L∞複雑性$|f_j|^\infty$などが挙げられる。ここで、f_jのL1複雑性、L2複雑性、L∞複雑性とは、関数（分離面）のノルム（大きさ）を表す量である。ベクトル$v=(v_1, \cdots, v_n)$に関して言えば、L1複雑性とは$\Sigma |v_i|$であり、L2複雑性とはΣv_i^2であり、L∞複雑性とは$\max |v_i|$となる。

[0051] 各クラス領域の大きさは、例えば図2に示されるクラス1の場合には超平面Aと超平面Bとで挟まれた領域の大きさ、例えば図3に示されるクラス3の場合には超球面Eと超球面Fとで挟まれた領域の大きさのことである。それらの大きさを表すために任意の基準を利用することが可能である。

[0052] 一般的には分離面の複雑性を大きくするほど学習データに対する分類誤差は小さくなるが、これは学習データに対して過学習をしているため未知の分類データに対する分類精度は低くなってしまう。したがって、分離面の複雑性を小さく保ったまま分類誤差を小さくする分離面を学習するために、両者の和（さらに各クラス領域の大きさの基準を加えた和）が最も小さくなる分離面集合を選択する。

[0053] 次に本実施の形態の動作を説明する。

[0054] 本実施の形態の動作は、分離面集合計算装置220による分離面の計算処理と、この計算された分離面を利用した分類対象データ150の分類処理とに大別される。

[0055] 分離面集合計算装置220による分離面の計算処理では、分離面集合最適化部221によって、学習データ記憶装置224から分類が既知の学習データを読み込み、この学習データに対する分類誤差の最小化、分離面集合の複雑性の最小化、および各クラス領域の大きさの最小化を同時に最適化する複数の分離面を計算して、記憶装置222に記憶する。次に、分離面集合出力部223によって、記憶装置222から複数の分離面を規定するデータを読み出し、分離面集合225として分離面集合記憶装置210に記憶する。

[0056] 本実施の形態のデータ分類装置200の動作は、図1に示した第1の実施の形態に関わるデータ分類装置100の動作と基本的に同じである。

[0057] このように本実施の形態によれば、第1の実施の形態と同様の効果が得られると同時に、分離面集合計算装置220によって計算した最新の複数の分離面で分離面集合記憶装置210に記憶された複数の分離面を置き換えることができ、学習データの充実にあわせて性能の向上を図ることができる効果がある。

[0058] [第3の実施の形態]

図9を参照すると、本発明の第3の実施の形態に関わるデータ分類装置300は、図7に示した第2の実施の形態に関わるデータ分類装置200と比較して、分離面集合記憶装置210に代えて超平面集合記憶装置310を有する点、および分離面集合計算装置220に代えて超平面集合計算装置320が接続されている点で相違する。

[0059] 超平面集合計算装置320は、1以上の既知クラスに分類されている複数の学習データおよびその分類に基づいて、特徴空間を1以上の既知クラスにそれぞれ対応する1以上のクラス領域とそれ以外の未知クラス領域とに分離する複数の超平面を計算する。1以上のクラス領域の各々は、この複数の分離面のうちの互いに交差しない2以上によって、他の領域と分離される。また、超平面集合記憶装置310は、超平面集合計算装置320で計算された複数の超平面を規定する情報を記憶する装置である。

[0060] 図10を参照すると、超平面集合計算装置320は、超平面集合最適化部321と、記憶装置222と、数理計画問題計算装置322と、超平面集合出力部323とを備える。超平面集合最適化部321は、学習データ記憶装置224から学習用のデータを入力する。超平面集合出力部323は、最適化された超平面集合324を出力する。すなわち、超平面集合計算装置320は、データ分類のために複数の互いに平行な超平面を計算する。従って、本実施の形態のデータ分類装置300では、図2に示されるように平行する超平面によって各クラスの領域を区切ることによってデータ分類を実現する。

[0061] 以下で、超平面の具体的な計算手順に関して、幾つかの例をもとに説明を行う。

[0062] 学習データ記憶装置224から入力されたデータに関するクラスのインデックスをj＝1, …, C(Cは1以上の整数)とする。以下では、x_i^jをj番目のクラスに属するi番目のデータとし、各クラスに属する学習データの数をN_jとする。特徴空間における超平面は、或る重みwおよび切片bに関して、$w^T \phi(x) + b = 0$を満たす点の集合として記述される。ここで、$f(x) = w^T \phi(x)$としておく。今、超平面は平行なため、重みwは共通であるから、wおよびj番目のクラスの超平面に対する切片b_j^+およびb_j^-が、超平面集合最適化部321によって最適化される。

[0063] なお、$\phi(x)$が線形の場合、特徴空間は、学習データ(および分類対象データ)と同じ次元数のベクトル空間になる。$\phi(x)$が非線形の場合、特徴空間は、学習データ(および分類対象データ)を非線形変換したベクトルデータと同じ次元数のベクトル空間になる。

[0064] 最適化のための基準として、以下の3条件、

(a)分類誤差最小化

(b)f(x)の複雑性最小化

(c)各既知クラス領域の大きさ最小化

を同時に最適化することによって、wおよび各jに対するb_j^+およびb_j^-を計算する。

[0065] 上記3条件に加えて、

(d)原点周囲の未知領域の大きさ最大化

(e)各クラスの領域が重ならない(あるいは各クラス領域の重なりの最小化)

の1つないしは双方をも同時に最適化することによってwおよび各jに対するb_j^+およびb_j^-を計算しても良い。

[0066] (c)の基準に関しては、超平面に対して各既知クラスの領域の大きさを最小化する。これによって、各クラス領域を両面からタイトに押さえることが要請される。

[0067] (d)の基準は各超平面に対して原点付近が未知クラスの領域になることを要請する。これは、学習データの張る空間の補空間のデータは未知クラスに属すると考えられるが、該データが学習データの張る空間へ射影された場合、必ず原点に射影されるためである。例として3次元の場合を考える。学習データが全てa(1, 0, 0)＋b(0, 1, 0)と表されるように、1次元目と2次元目のみに分布していると仮定する。この場合、3次元目に分布する未知クラスのデータc(0, 0, 1)は1次元目と2次元目の成分が0のため、データの張る空間に対しては必ず原点に射影される。

[0068] (a)から(e)の複数の基準を同時に最適化する具体的な例を幾つか挙げる。

[0069] [C＝1の場合]

学習データ記憶装置224から入力されたデータに関するクラスが唯一の場合には、互いに平行な2つの超平面が計算される。そのような2つの超平面は、例として(2)式に示される最適化問題を解くことで求められる。

[0070]

$$\min \quad \frac{1}{2}w'w+\frac{1}{N_1}\sum_i\left(\xi_i^{1+}+\xi_i^{1-}\right)+\nu_1\left(b_1^+-b_1^-\right)-\nu_0\left(b_1^++b_1^-\right)$$

$$subject \quad to$$

$$w'\phi\left(x_i^1\right)-b_1^+\le\xi_i^{1+} \qquad w'\phi\left(x_i^1\right)-b_1^-\ge-\xi_i^{1-}$$

$$b_1^+\ge b_1^- \quad b_1^-\ge 0 \quad \xi_i^{1+}\ge 0 \quad \xi_i^{1-}\ge 0 \qquad \cdots(2)$$

[0071] (2)式では(a)から(d)の基準が、(a)第2項、(b)第1項、(c)第3項、(d)第4項として表現されている。(e)の基準に関しては1クラスの場合には考慮される必要がない。ν_0およびν_1はどの基準に重きをおくかを決定するパラメータで、0より大きく1より小さい実数値である。(2)式によって計算される2つの超平面は、図11に示されるような超平面となる。以下、(2)式における目的関数および制約条件について説明する。

[0072] 式(2)の目的関数における第1項は、最適化の基準(b)のために必要な項であり、

複雑性としてL2複雑性を採用すると、f(x)のL2複雑性はこのように計算される。第2項は、最適化の基準(a)のために必要な項であり、ξ_i^{1+}とξ_i^{1-}は誤差を表すためのスラック変数である。第3項は、最適化の基準(c)のために必要な項であり、$b_1^- \leqq w'\phi(x_i^1) \leqq b_1^+$のため、$b_1^- - b_1^+$を小さくすることによって既知クラスを包む領域を最小化している。第4項は、最適化の基準(d)のために必要な項である。原点周辺の未知領域の大きさを最大化することは、すなわち既知領域を原点から遠ざけることを意味する。そのため、既知領域の中心$(b_1^- + b_1^+)/2$を原点から遠ざけることで(d)の基準が達成される。

[0073] 式(2)の制約条件における、$w'\phi(x_i^1) - b_1^+ \leqq \xi_i^{1+}$、$w'\phi(x_i^1) - b_1^- \geqq -\xi_i^{1-}$、$\xi_i^{1+} \geqq 0$、$\xi_i^{1-} \geqq 0$は、次の意味を持つ。すなわち、図11に示されるように、クラス1に属するデータはb_1^+とb_1^-の間に入る必要がある(つまり、$b_1^- \leqq w'\phi(x_i^1) \leqq b_1^+$である)が、入らなかった分は誤差としてカウントする。$b_1^+ \geqq b_1^-$は、$b_1^- \leqq w'\phi(x_i^1) \leqq b_1^+$のために必要な制約条件である。$b_1^- \geqq 0$は、原点領域を未知領域とするために必要な制約条件である。つまり、$b_1^- \geqq 0$という制約条件がないと、$b_1^- \leqq 0 \leqq b_1^+$となり得るためである。なお、$b_1^- \geqq 0$の代わりに、$b_1^+ \leqq 0$でも良い。

[0074] (2)式は標準的な凸2次計画問題であり、超平面集合最適化部321および数理計画問題計算装置322によって最適解が計算される。

[0075] なお、特徴空間が非線形で、特徴空間への写像φが明示的に与えられていない場合には、一般には(2)式を直接解くことができない。しかし、特徴空間における内積がカーネル関数として定義されている場合には、(2)式の双対問題を解くことで超平面が計算される。

[0076] (2)式の双対問題は(3)式のようにラグランジュの未定乗数を導入することによって、(4)式となる。

[0077]

$$L_P = \frac{1}{2}w'w + \frac{1}{N_1}\sum_i\left(\xi_i^{1+} + \xi_i^{1-}\right) + \nu_1\left(b_1^+ - b_1^-\right) - \nu_0\left(b_1^+ + b_1^-\right)$$
$$- \sum_i \alpha_i^{1+}\left(\xi_i^{1+} - w'\phi(x_i^1) + b_1^+\right) - \sum_i \alpha_i^{1-}\left(\xi_i^{1-} + w'\phi(x_i^1) - b_1^-\right)$$
$$- \mu_1\left(b_1^+ - b_1^-\right) - \mu_0 b_1^- - \sum_i\left(\gamma_i^{1+}\xi_i^{1+} - \gamma_i^{1-}\xi_i^{1-}\right) \qquad \cdots(3)$$

[0078]

$$\max \quad -\frac{1}{2}\sum_{i,i'}\left(\alpha_i^{1-}-\alpha_i^{1+}\right)\left(\alpha_{i'}^{1-}-\alpha_{i'}^{1+}\right)k\left(x_i^1,x_{i'}^1\right)$$

$$subject \quad to$$

$$\sum_i \alpha_i^{1+}+\mu_1=\nu_1 \quad \sum_i \alpha_i^{1-}+\mu_1-\mu_0=\nu_1$$

$$0\le\alpha_i^{1+}\le\frac{1}{N_1} \quad 0\le\alpha_i^{1-}\le\frac{1}{N_i} \quad \mu_1,\mu_0\ge 0 \qquad \cdots(4)$$

[0079] ラグランジュの未定乗数は、α_i^{1+}, α_i^{1-}, μ_0, μ_1, γ_i^{1+}, γ_i^{1-}, δ である。ただし、$k(x_i^1, x_i'^1)=\phi(x_i^1)^T\phi(x_i'^1)$は特徴空間での内積であり、双対問題では$\phi(x)$がどのような関数であっても、その内積$\phi(x_i^1)^T\phi(x_i'^1)$さえ計算できれば解くことが可能である。(4)式で表される双対問題も凸2次計画問題である。

[0080] 双対問題に対して重みwは、(5)式と表されるため、$f(x)=w^T\phi(x)$は(6)式で表される。相対問題を解いた場合には、記憶する内容が図4のw_iとbの組でなくて、α_iとbの組になる。

[0081]

$$w=\sum_i\left(\alpha_i^{1-}-\alpha_i^{1+}\right)\phi\left(x_i^1\right) \qquad \cdots(5)$$

[0082]

$$f(x)=\sum_i\left(\alpha_i^{1-}-\alpha_i^{1+}\right)\phi\left(x_i^1\right)\phi(x)=\sum_i\left(\alpha_i^{1-}-\alpha_i^{1+}\right)k\left(x_i^1,x\right) \qquad \cdots(6)$$

[0083] [C＝2の場合]

学習データ記憶装置224から入力されたデータに関するクラスが2つの場合、各クラスに対して平行な2つの超平面が計算される。そのような複数の超平面は、例として(7)式に示される最適化問題を解くことで計算される。

[0084]

$$\min \quad \frac{1}{2}w'w+\frac{1}{N}\sum_{i,j}\left(\xi_i^{j+}+\xi_i^{j-}\right)+\nu_1\sum_j\left(b_j^+-b_j^-\right)-\nu_0\sum_j\left|b_j^++b_j^-\right|$$

$$subject \quad to$$

$$w'\phi\left(x_i^j\right)-b_j^+\le\xi_i^{j+} \qquad w'\phi\left(x_i^j\right)-b_j^-\ge-\xi_i^{j-} \quad b_j^+\ge b_j^-$$

$$b_1^-\ge 0 \quad 0\ge b_2^+ \quad \xi_i^{j+}\ge 0 \quad \xi_i^{j-}\ge 0 \qquad \cdots(7)$$

[0085] (7)式では(a)から(e)の基準が、(a)第2項、(b)第1項、(c)第3項、(d)第4項として表現されている。(e)の基準に関しては$b_1^- \geqq b_2^+$が自動的に満たされるため、明示的に考慮する必要はない。ν_0およびν_1およびν_2はどの基準に重きをおくかを決定するパラメータで、0より大きく1より小さい実数値である。(7)式によって計算される複数の超平面は、図12に示されるような超平面となる。以下、(7)式における目的関数および制約条件について説明を補足する。

[0086] 式(7)の目的関数における第4項は、最適化の基準(e)のために必要な項であり、絶対値記号が付いている理由は、j=2はb_2^-, b_2^+ともに負になるためである。式(7)の制約条件における$0 \geqq b_2^+$は、2つあるクラスの双方とも原点0を跨がないようにするための制約条件である。つまり、$b_1^- \leqq 0 \leqq b_1^+$や$b_2^- \leqq 0 \leqq b_2^+$という状況を避けるためには、次の3通りが考えられる。両方のクラスが正側(つまり、$0 \leqq b_1^-$かつ$0 \leqq b_2^-$)、両方のクラスが負側(つまり、$b_1^+ \leqq 0$かつ$b_2^+ \leqq 0$)、各クラスが原点0を挟んで互いに反対側。式(7)では最後のケースを採用している。

[0087] C=1の場合と同様に、(7)式は凸2次計画問題である。また、(2)式から(4)式を得たのと同様の手順で双対問題を導出し、双対問題を解くことで最適化を行うことも可能である。(7)式の双対問題も凸2次計画問題となる。

[0088] [C≧3の場合]

学習データ記憶装置224から入力されたデータに関するクラスが3以上の一般の場合に、平行な複数の超平面の組を計算するには、入力されたクラスの任意の2つの組み合わせに対してC=2の場合の最適化を実施し、得られた複数の超平面の組を利用して多数決をとることが考えられる。

[0089] また例えば(8)式に示される最適化問題を解くことで、平行な複数の超平面の組を計算することもできる。

[0090]

$$\min \quad \frac{1}{2}w'w+\frac{1}{N}\sum_{i,j}\left(\xi_i^{j+}+\xi_i^{j-}\right)+\sum_j \nu_j\left(b_j^+-b_j^-\right)-\nu_0\left(b_0^+-b_0^-\right)$$

$$subject \quad to$$

$$w'\phi\left(x_i^j\right)-b_j^+\le\xi_i^{j+} \quad w'\phi\left(x_i^j\right)-b_j^-\ge-\xi_i^{j-} \quad b_j^+\ge b_j^-$$

$$b_0^+\ge 0 \quad 0\ge b_0^- \quad \xi_i^{j+}\ge 0 \quad \xi_i^{j-}\ge 0$$

$$b_j^-\ge b_k^+-\psi_{jk}^- \quad b_j^+\le b_k^-+\psi_{jk}^+ \quad \psi_{jk}^-\psi_{jk}^+=0$$

$$b_j^-\ge b_0^+-\psi_{j0}^- \quad b_0^-\le b_j^++\psi_{j0}^+ \quad \psi_{j0}^-\psi_{j0}^+=0 \qquad \cdots(8)$$

[0091] (8)式では(a)から(e)の基準が、(a)第2項、(b)第1項、(c)第3項、(d)第4項として表現されている。(e)の基準に関してはφに関する制約条件によって表現されている。以下、(8)式における目的関数および制約条件について説明を補足する。

[0092] 式(2)および式(7)で示した1クラスおよび2クラスの場合には、特徴空間におけるクラスの領域の順序が決まっていたため、各クラスの領域を原点から遠ざけるという方式で、基準(e)を達成することができた。しかし、一般に多クラスになるとクラスの領域の順序をどのようにすれば良いかは自明でない。一つの案として、それを全ての組み合わせで解く方法が考えられるが、計算量が多くなる欠点がある。(8)式による最適化は、組み合わせを考えること無しに、自動的に順序が最も良いものに決定されるようにしている。

[0093] そのために、まず、図13に示されるように、原点周りの未知クラスの領域がb_0^-とb_0^+に挟まれた領域とし、そのための制約条件として$b_0^+\geqq 0$、$0\geqq b_0^-$を置き、目的関数の第4項で、その領域を最大化している(第4項の符号はマイナスで、目的関数が最小化だから最大化になる)。

[0094] 次に、既知クラスの領域(および原点周りの未知クラス領域)が図14に示すようにオーバーラップしてはいけないための制約が必要になる。このような制約は、各クラスの領域の順序と原点との位置関係が明示的に決まっている場合には、$b_1^-\leqq 0$、$b_2^-\geqq 0$、$b_2^+\leqq b_3^-$というように明示的に順序として重複しない制約を書くことが可能である。全組み合わせを考える場合にはこのような制約条件をつけるが、(8)式は順序が事前にわからないことを前提としているため、そのような制約は書けない。そこで、$b_j^-\geqq$

$b_k^+ - \Psi_{jk}^-$、$b_j^+ \leqq b_k^- + \Psi_{jk}^+$、$\Psi_{jk}^- \Psi_{jk}^+ = 0$および、$b_j^- \geqq b_0^+ - \Psi_{j0}^-$、$b_0^- \leqq b_j^+ + \Psi_{j0}^+$、$\Psi_{j0}^- \Psi_{j0}^+ = 0$という制約条件により、既知クラスの領域（および原点周りの未知クラス領域）がオーバーラップしてはいけないための制約を課している。

[0095] また、$b_j^- \geqq b_k^+ - \Psi_{jk}^-$に関して、$b_j^- \geqq b_k^+$が成り立つ場合（つまり、クラスjがクラスkより正の方向にある）には、$\Psi_{jk}^- = 0$になる。逆に、$b_j^+ \leqq b_k^- + \Psi_{jk}^+$に関して、$b_j^+ \leqq b_k^-$が成り立つ場合（つまり、クラスjがクラスkより負の方向にある）には、$\Psi_{jk}^+ = 0$になる。クラス間の重複がないためには、$b_j^- \geqq b_k^+$または$b_j^+ \leqq b_k^-$とならなくてはならないので、$\Psi_{jk}^- = 0$または$\Psi_{jk}^+ = 0$が成り立つ必要がある。したがって、$\Psi_{jk}^- \Psi_{jk}^+ = 0$という制約によって、各クラスの重複がないという制約を課すことが可能である。

[0096] Ψ_{j0}^+, Ψ_{j0}^-に関する制約条件は、原点周りの領域と既知クラスの領域に関する同様の制約を表す。

[0097] 次に本実施の形態の動作を説明する。

[0098] 本実施の形態の動作は、超平面集合計算装置320による超平面の計算処理と、この計算された超平面を利用した分類対象データ150の分類処理とに大別される。

[0099] 超平面集合計算装置320による超平面の計算処理では、超平面集合最適化部321によって、学習データ記憶装置224から分類が既知の学習データを読み込み、この学習データに対する分類誤差の最小化、超平面集合の複雑性の最小化、および各クラス領域の大きさの最小化を同時に最適化する複数の超平面を計算して、記憶装置222に記憶する。次に、超平面集合出力部323によって、記憶装置222から複数の超平面を規定するデータを読み出し、超平面集合324として超平面集合記憶装置310に記憶する。

[0100] 本実施の形態のデータ分類装置300の動作は、図1に示した第1の実施の形態に関わるデータ分類装置100の動作と基本的に同じである。

[0101] このように本実施の形態によれば、第1の実施の形態と同様の効果が得られると同時に、超平面集合計算装置320によって計算した最新の複数の超平面で超平面集合記憶装置310に記憶された複数の超平面を置き換えることができ、学習データの充実にあわせて性能の向上を図ることができる効果がある。

[0102] ［第4の実施の形態］

図15を参照すると、本発明の第4の実施の形態に関わるデータ分類装置400は、図7に示した第2の実施の形態に関わるデータ分類装置200と比較して、分離面集合記憶装置210に代えて超球面集合記憶装置410を有する点、および分離面集合計算装置220に代えて超球面集合計算装置420が接続されている点で相違する。

[0103] 超球面集合計算装置420は、1以上の既知クラスに分類されている複数の学習データおよびその分類に基づいて、特徴空間を1以上の既知クラスにそれぞれ対応する1以上のクラス領域とそれ以外の未知クラス領域とに分離する複数の超球面であって、1クラス当たり2個以上で且つ互いに同心の複数の超球面を計算する装置である。また、超球面集合記憶装置410は、超球面集合計算装置420で計算された複数の超球面を規定する情報を記憶する装置である。

[0104] 図16を参照すると、超球面集合計算装置420は、超球面集合最適化部421と、記憶装置222と、数理計画問題計算装置422と、超球面集合出力部423とを備え、学習データ記憶装置224から学習用のデータを入力し、最適化された超球面集合424を出力する。すなわち、超球面集合計算装置420は、データ分類のために複数の同心の超球面を計算する。従って、本実施の形態のデータ分類装置400では、図3に示されるように同心の超球面によって各クラスの領域を区切ることによってデータ分類を実現する。

[0105] 以下で、超球面の具体的な計算手順に関して、幾つかの例をもとに説明を行う。

[0106] 学習データ記憶装置224から入力されたデータに関するクラスのインデックスをj＝1, …, Cとする。以下では、x_i^jをj番目のクラスに属するi番目のデータとし、各クラスに属する学習データの数をN_jとする。超球面の中心をc、半径をrとすると、超球面は$|\phi(x)-c|^2=r$と書ける。今、超球面は同心であるため、中心cは各クラスで共通であるから、cおよびj番目のクラスに対する外側の半径r_j^+および内側の半径r_j^-が超球面集合最適化部421によって最適化される。

[0107] 最適化のための基準としては、以下の3つの条件を同時に最適化することによって、cおよび各jに対するr_j^+およびr_j^-を計算する。

(a')分類誤差最小化

(b')cの複雑性最小化

(c')各既知クラス領域の大きさ最小化。

[0108] なお、上記以外にさらに、

(d')原点周囲の未知領域の大きさ最大化

(e')各クラスの領域が重ならない

の1つないしは双方をも同時に最適化することによってcおよび各jに対するr_j^+およびr_j^-を計算しても良い。

[0109] (a')から(e')の複数の基準を同時に最適化する具体的な例は、例えば式(9)が挙げられる。式(9)はクラスが幾つでも適用可能であるが、クラスの順序が判っていることが前提となっている。

[0110]

$$\min \quad \sum_j \left(r_j^{+2} - r_j^{-2}\right) + \frac{1}{N}\sum_{i,j}\left(\xi_i^{j+} + \xi_i^{i-}\right) - \nu_0 \min\left\{r_j^-\right\}$$

$$subject \quad to$$

$$\left|\phi\left(x_i^j\right) - c\right|^2 \le r_j^{+2} + \xi_i^{j+} \quad \left|\phi\left(x_i^j\right) - c\right|^2 \ge r_j^{-2} - \xi_i^{j-}$$

$$r_j^+ \ge r_j^- \quad r_{j+1}^- \ge r_j^+ \quad \min\left\{r_j^-\right\} \ge 0 \quad c^2 \le \min\left\{r_j^-\right\}^2$$

$$\xi_i^{j+} \ge 0 \quad \xi_i^{j-} \ge 0 \qquad \cdots(9)$$

[0111] 式(9)によって計算される超球面集合の一例を図17に示す。式(9)は目的関数および制約条件の凹部分と凸部分が加算の形になっているため、Concave－Convex Procedure(文献8参照)などを利用して効率的に最適解を計算することが可能である。以下、式(9)における目的関数および制約条件について説明する。

[0112] 式(9)の目的関数における第1項は、クラスjの領域の外半径－内半径という形なので、最適化の基準(c')のために必要な項である。第2項は、式(7)の第2項に相当し、最適化の基準(a')のために必要な項である。第3項は、最適化の基準(d')のために必要な項である。その理由は次の通りである。

[0113] まず、制約条件の$c^2 \leqq \min\{r_j^-\}^2$から、原点は一番小さい超球面の内側にあることが制約として課される。c^2は原点と超球面の中心との距離で、$\min\{r_j^-\}^2$は超球面の中心と一番内側の超球面との距離(つまり半径)であるためである。つまり、一番内側の球の内部が、原点周りの未知領域になる。したがって、$\min\{r_j^-\}^2$を大きくすること

によって、基準(d')が達成される。

[0114] 基準(b')は式(9)の目的関数の中では明示的には含まれておらず、制約条件の中に暗黙的に含まれている。基準(e')は、$r_j^+ \geqq r_j^-$と$r_{j+1}^- \geqq r_j^+$によって制約されている。

[0115] 次に本実施の形態の動作を説明する。

[0116] 本実施の形態の動作は、超球面集合計算装置420による超球面の計算処理と、この計算された超球面を利用した分類対象データ150の分類処理とに大別される。

[0117] 超球面集合計算装置420による超球面の計算処理では、超球面集合最適化部421が、学習データ記憶装置224から分類が既知の学習データを読み込み、この学習データに対する分類誤差の最小化、超球面集合の複雑性の最小化、および各クラス領域の大きさの最小化を同時に最適化する複数の超球面を計算して、記憶装置222に記憶する。次に、超球面集合出力部323が、記憶装置222から複数の超球面を規定するデータを読み出し、超球面集合424として超球面集合記憶装置410に記憶する。

[0118] 本実施の形態のデータ分類装置400の動作は、図1に示した第1の実施の形態に関わるデータ分類装置100の動作と基本的に同じである。

[0119] このように本実施の形態によれば、第1の実施の形態と同様の効果が得られると同時に、超球面集合計算装置420によって計算した最新の複数の超球面で超球面集合記憶装置410に記憶された複数の超球面を置き換えることができる。そのため、学習データの充実にあわせて性能の向上を図ることができる効果がある。

[0120] 以上本発明の実施の形態について説明したが、本発明は以上の実施の形態にのみ限定されず、その他各種の付加変更が可能である。また、本発明のデータ分類装置は、その有する機能をハードウェア的に実現することは勿論、コンピュータとプログラムとで実現することができる。プログラムは、磁気ディスクや半導体メモリ等のコンピュータ可読記録媒体に記録されて提供され、コンピュータの立ち上げ時などにコンピュータに読み取られ、そのコンピュータの動作を制御することにより、そのコンピュータを前述した各実施の形態におけるデータ分類装置、分離面集合計算装置、超平面集合計算装置、超球面集合計算装置として機能させ、前述した処理を行わせる。

請求の範囲

[1] 特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を規定する情報を記憶する分離面集合記憶部と、前記少なくとも1つの既知クラス領域の各々は前記複数の分離面のうちの互いに交差しない2以上によって外部領域と分離され、

内積が計算可能な分類対象データが、前記分離面記憶部に記憶された前記情報で規定される前記少なくとも1つの既知クラス領域と前記未知クラス領域とのうちのどの領域に属するかを計算することによって、前記分類対象データの分類を決定する分類部

とを具備するデータ分類装置。

[2] 前記特徴空間における内積が計算可能であり前記少なくとも1つの既知クラスのいずれかにそれぞれ分類されている複数の学習データおよび前記複数の学習データの各々の分類に基づいて複数の分離面を計算し、前記複数の分離面を規定する情報を前記分離面集合記憶部に記憶する分離面集合計算部

を更に具備する請求の範囲1に記載のデータ分類装置。

[3] 前記分離面集合計算部は、前記複数の学習データに対する分類誤差の最小化、前記複数の分離面の各々の複雑性の最小化、および前記少なくとも1つの既知クラス領域の大きさの最小化をそれぞれ最適化目的として前記複数の分離面を計算する請求の範囲2に記載のデータ分類装置。

[4] 前記分離面集合計算部は、原点周囲の前記未知クラス領域の大きさの最大化をさらに最適化目的の一つとする請求の範囲3に記載のデータ分類装置。

[5] 前記分離面集合計算部は、前記少なくとも1つの既知クラス領域の相互間の重なりの最小化をさらに最適化目的の一つとする請求の範囲3に記載のデータ分類装置。

[6] 前記複数の分離面の各々が、前記特徴空間上で開いた超平面を成す請求の範囲1乃至5の何れか1項に記載のデータ分類装置。

[7] 前記複数の分離面の各々が、前記特徴空間上で閉じた超平面を成す請求の範囲1乃至5の何れか1項に記載のデータ分類装置。

[8] 前記特徴空間が、前記学習データおよび前記分類対象データと同じ次元数のベク

トル空間である請求の範囲1乃至7の何れか1項に記載のデータ分類装置。

[9] 前記特徴空間が、前記学習データおよび前記分類対象データに対する非線形変換によって特徴付けられる空間である請求の範囲1乃至7の何れか1項に記載のデータ分類装置。

[10] (a)特徴空間における内積が計算可能な分類対象データを入力する工程と、

(b)特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を分離面記憶部から入力する工程と、前記少なくとも1つの複数の既知クラス領域の各々は前記複数の分離面のうちの互いに交差しない2以上によって外部領域と分離され、

(c)前記分類対象データが、少なくとも1つの既知クラス領域と前記未知クラス領域とのうちのどの領域に属するかを計算することによって、前記分類対象データの分類を決定する工程

とを具備するデータ分類方法。

[11] (d)前記特徴空間における内積が計算可能であり前記少なくとも1つの既知クラスのいずれかにそれぞれ分類されている複数の学習データおよび前記複数の学習データの各々の分類とに基づいて前記複数の分離面を計算し、前記複数の分離面を規定する情報を前記分離面集合記憶部に記憶する工程

を更に具備する請求の範囲10に記載のデータ分類方法。

[12] 前記工程(d)では、前記複数の学習データに対する分類誤差の最小化、前記複数の分離面の各々の複雑性の最小化、および前記少なくとも1つの既知クラス領域の最小化をそれぞれ最適化目的として前記複数の分離面を計算する請求の範囲11に記載のデータ分類方法。

[13] 前記工程(d)では、原点周囲の前記未知クラス領域の大きさの最大化をさらに最適化目的の一つとする請求の範囲12に記載のデータ分類方法。

[14] 前記工程(d)では、前記少なくとも1つの既知クラス領域の相互間の重なりの最小化をさらに最適化目的の一つとする請求の範囲12に記載のデータ分類方法。

[15] 特徴空間における内積が計算可能であり少なくとも1つの既知クラスのいずれかにそれぞれ分類されている複数の学習データを記憶する学習データ記憶部と、

前記学習データ記憶部に記憶された前記複数の学習データおよび前記複数の学習データの各々の分類に基づいて、前記特徴空間を前記少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を計算する分離面集合計算部と、前記少なくとも1つの既知クラス領域の各々は前記複数の分離面のうちの互いに交差しない2以上によって外部領域と分離され、

前記複数の分離面を規定する情報を記憶する分離面集合記憶部

とを具備する分離面集合計算装置。

[16] 前記分離面集合計算部は、前記複数の学習データに対する分類誤差の最小化、前記複数の分離面の各々の複雑性の最小化、および前記少なくとも1つの既知クラス領域の大きさの最小化をそれぞれ最適化目的として前記複数の分離面を計算する請求の範囲15に記載の分離面集合計算装置。

[17] 前記分離面集合計算部は、原点周囲の前記未知クラス領域の大きさの最大化をさらに最適化目的の一つとする請求の範囲16に記載の分離面集合計算装置。

[18] 前記分離面集合計算部は、前記少なくとも1つの既知クラス領域の相互間の重なりの最小化をさらに最適化目的の一つとする請求の範囲16に記載の分離面集合計算装置。

[19] (a)特徴空間における内積が計算可能な分類対象データを入力する工程と、

(b)特徴空間を少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を分離面記憶部から入力する工程と、前記少なくとも1つの複数の既知クラス領域の各々は前記複数の分離面のうちの互いに交差しない2以上によって外部領域と分離され、

(c)前記分類対象データが、少なくとも1つの既知クラス領域と前記未知クラス領域とのうちのどの領域に属するかを計算することによって、前記分類対象データの分類を決定する工程

とを具備する方法をコンピュータに実行させるためのプログラム。

[20] (d)前記特徴空間における内積が計算可能であり前記少なくとも1つの既知クラスのいずれかにそれぞれ分類されている複数の学習データおよび前記複数の学習デー

タの各々の分類とに基づいて前記複数の分離面を計算し、前記複数の分離面を規定する情報を前記分離面集合記憶部に記憶する工程

を更に具備する方法をコンピュータに実行させるための請求の範囲19に記載のプログラム。

[21] 前記工程(d)では、前記複数の学習データに対する分類誤差の最小化、前記複数の分離面の各々の複雑性の最小化、および前記少なくとも1つの既知クラス領域の最小化をそれぞれ最適化目的として前記複数の分離面を計算する請求の範囲20に記載のプログラム。

[22] 前記工程(d)では、原点周囲の前記未知クラス領域の大きさの最大化をさらに最適化目的の一つとする請求の範囲21に記載のプログラム。

[23] 前記工程(d)では、前記少なくとも1つの既知クラス領域の相互間の重なりの最小化をさらに最適化目的の一つとする請求の範囲21に記載のプログラム。

[24] (a)特徴空間における内積が計算可能であり少なくとも1つの既知クラスのいずれかにそれぞれ分類されている複数の学習データを記憶する工程と、

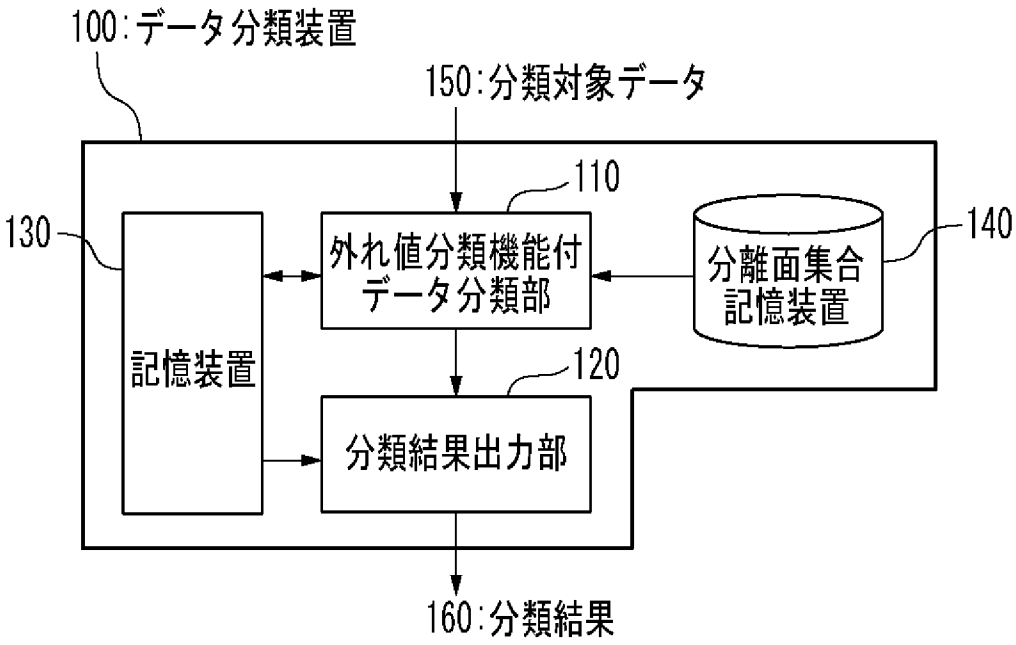
(b)前記学習データ記憶部に記憶された前記複数の学習データおよび前記複数の学習データの各々の分類に基づいて、前記特徴空間を前記少なくとも1つの既知クラスにそれぞれ対応する少なくとも1つの既知クラス領域と未知クラス領域とに分離する複数の分離面を計算する工程と、前記少なくとも1つの既知クラス領域の各々は前記複数の分離面のうちの互いに交差しない2以上によって外部領域と分離され、

(c)前記複数の分離面を規定する情報を記憶する工程

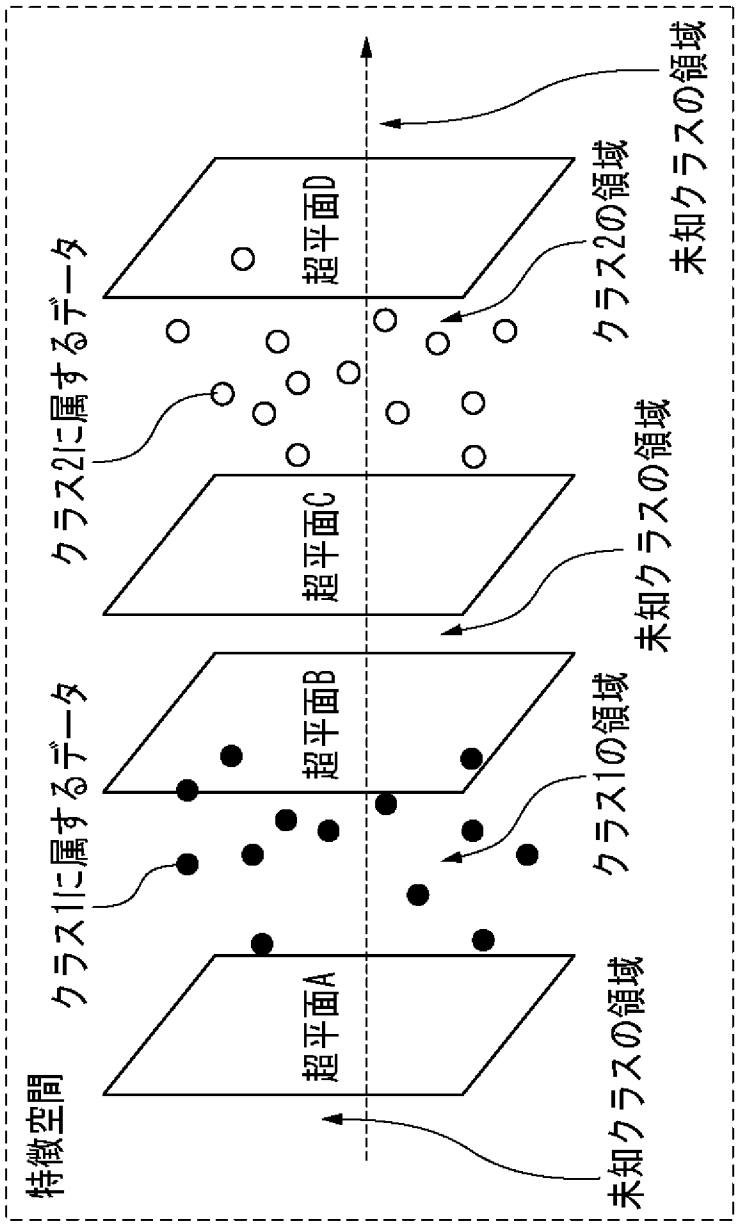
とを具備する方法をコンピュータに実行させるためのプログラム。

[25] 前記(c)計算する工程は、前記分離面集合計算部は、前記複数の学習データに対する分類誤差の最小化、前記複数の分離面の各々の複雑性の最小化、および前記少なくとも1つの既知クラス領域の大きさの最小化をそれぞれ最適化目的として前記複数の分離面を計算する請求の範囲24に記載のプログラム。

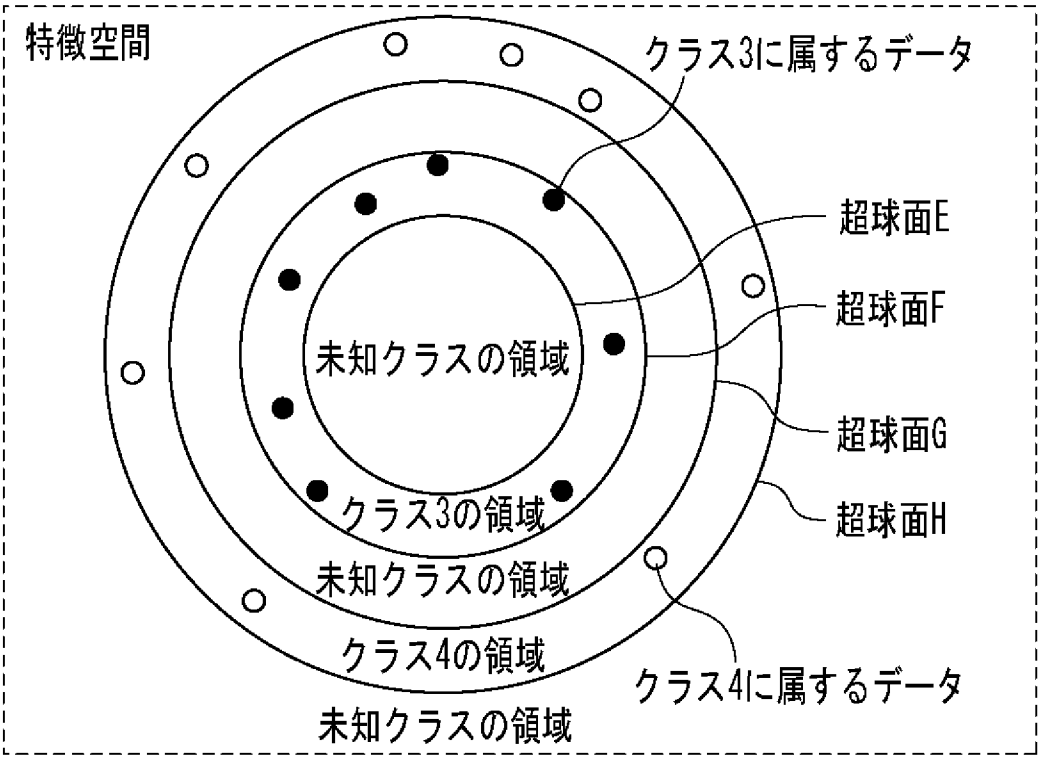
[図1]


100:データ分類装置
150:分類対象データ
110
130
外れ値分類機能付
データ分類部
分離面集合
記憶装置
140
記憶装置
120
分類結果出力部
160:分類結果

[図2]


特徴空間
クラス1に属するデータ
クラス2に属するデータ
超平面A
超平面B
超平面C
超平面D
未知クラスの領域
クラス1の領域
未知クラスの領域
クラス2の領域
未知クラスの領域

[図3]



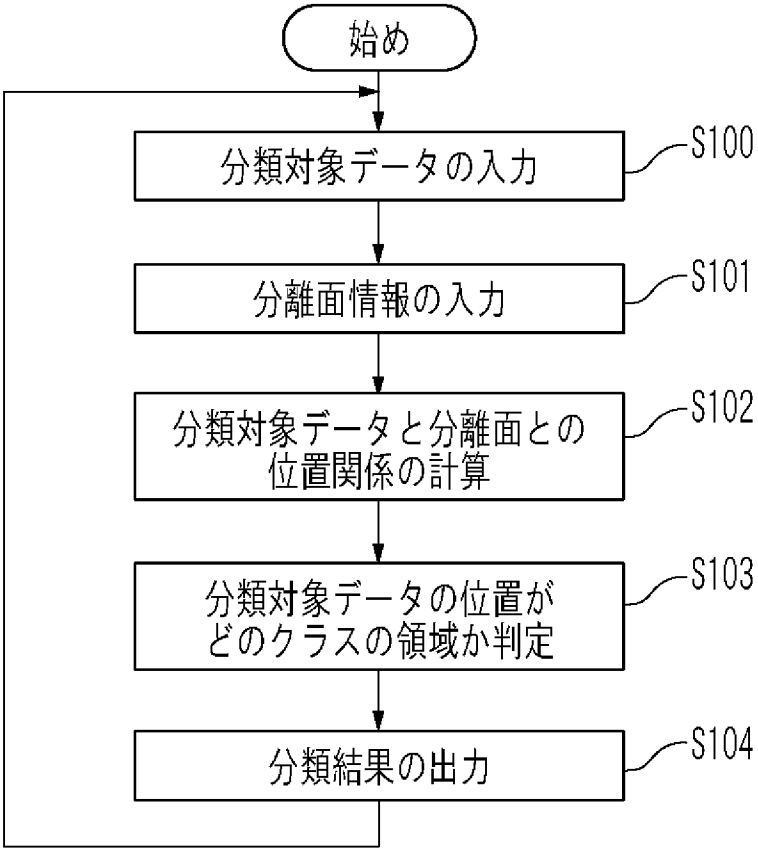
[図4]

超平面番号	w_1	…	w_M	b
超平面A	w_1^A	…	w_M^A	b^A
超平面B	w_1^B	…	w_M^B	b^B
超平面C	w_1^C	…	w_M^C	b^C
超平面D	w_1^D	…	w_M^D	b^D

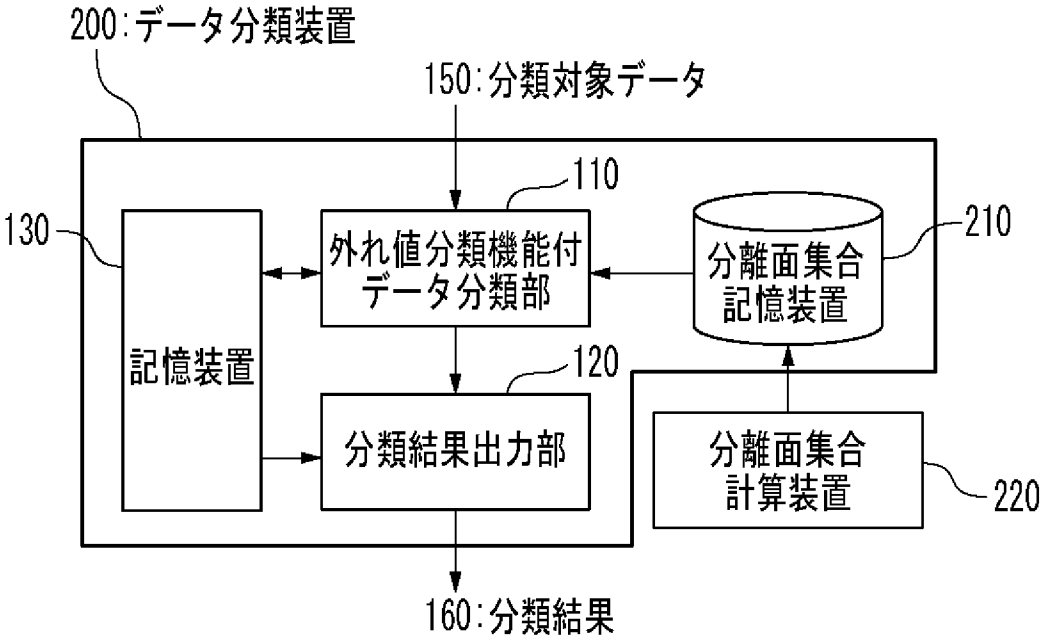
[図5]

超球面番号	w_1	…	w_M	r
超球面E	w_1^E	…	w_M^E	r^E
超球面F	w_1^F	…	w_M^F	r^F
超球面G	w_1^G	…	w_M^G	r^G
超球面H	w_1^H	…	w_M^H	r^H

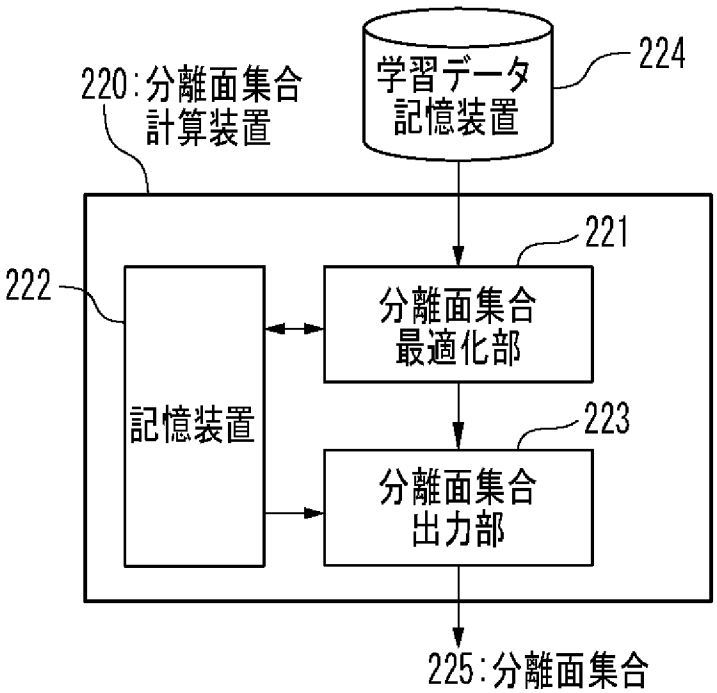
[図6]



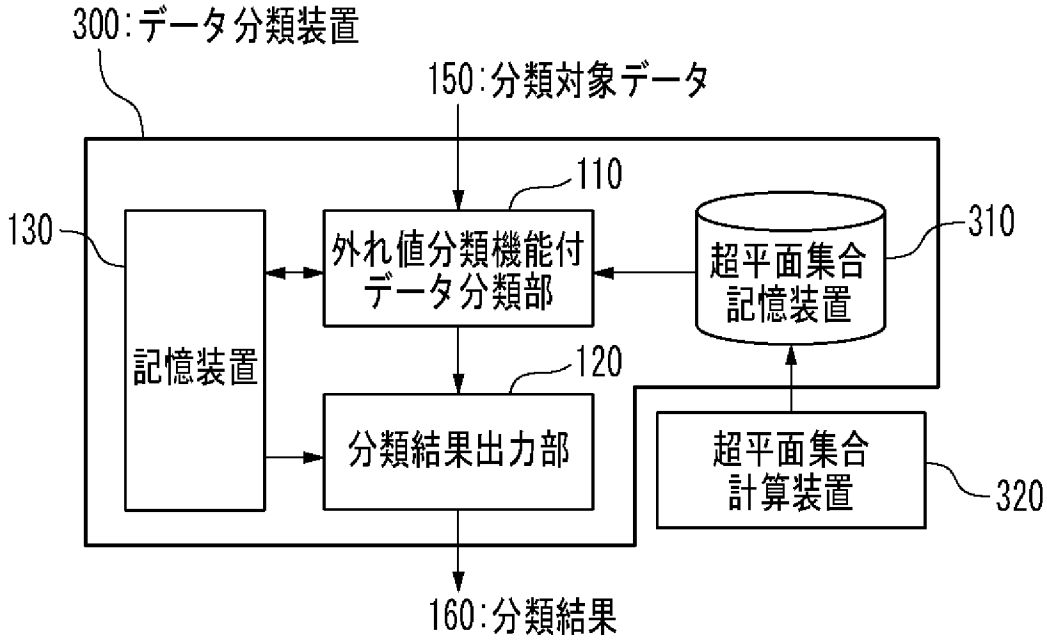
[図7]


200:データ分類装置
150:分類対象データ
110
130
外れ値分類機能付
データ分類部
分離面集合
記憶装置
210
記憶装置
120
分類結果出力部
分離面集合
計算装置
220
160:分類結果

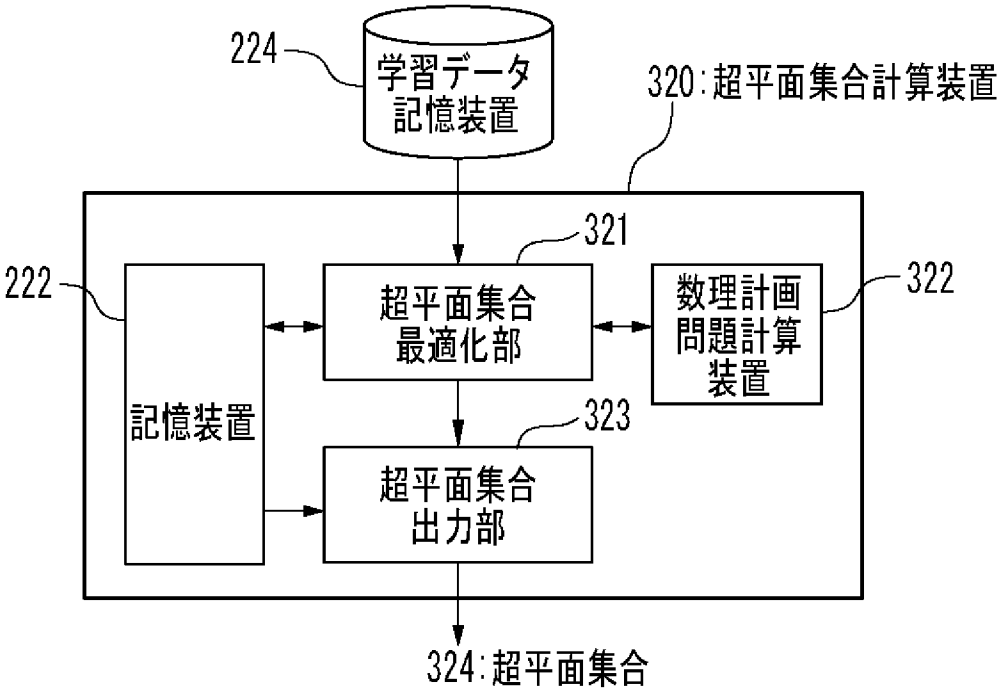
[図8]


220:分離面集合
計算装置
学習データ
記憶装置
224
221
222
分離面集合
最適化部
記憶装置
223
分離面集合
出力部
225:分離面集合

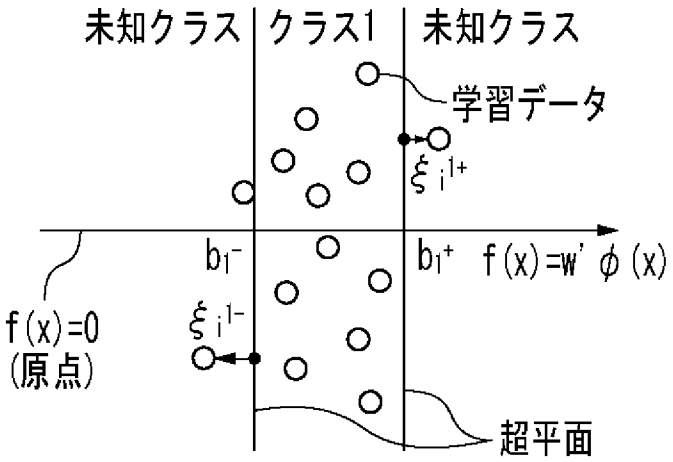
[図9]


300:データ分類装置
150:分類対象データ
110
130
外れ値分類機能付
データ分類部
超平面集合
記憶装置
310
記憶装置
120
分類結果出力部
超平面集合
計算装置
320
160:分類結果

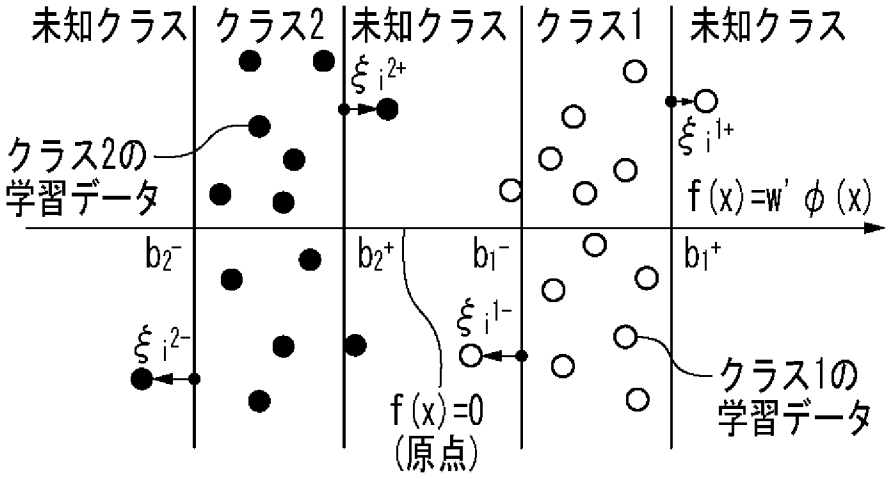
[図10]


224
学習データ
記憶装置
320:超平面集合計算装置
321
222
超平面集合
最適化部
数理計画
問題計算
装置
322
記憶装置
323
超平面集合
出力部
324:超平面集合

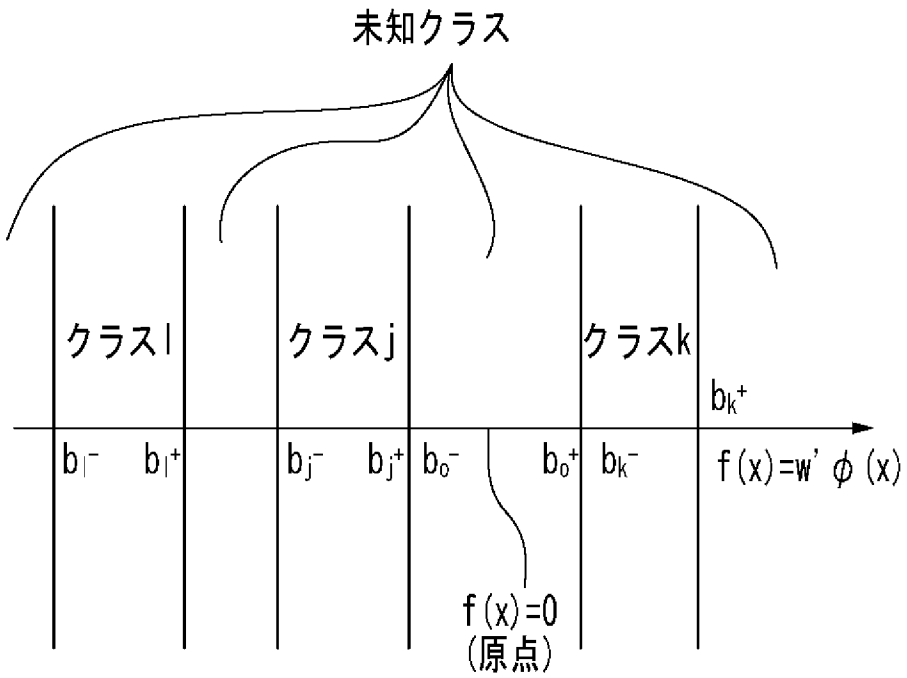
[図11]


未知クラス
クラス1
未知クラス
学習データ
ξ_i^{1+}
b_1^-
b_1^+
$f(x)=w'\phi(x)$
$f(x)=0$
(原点)
ξ_i^{1-}
超平面

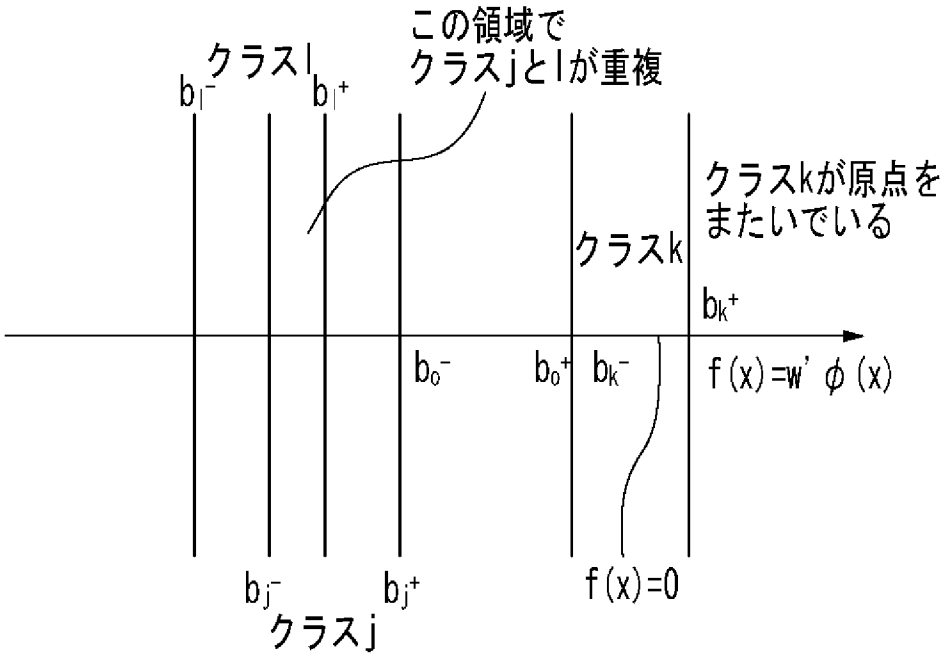
[図12]


未知クラス
クラス2
未知クラス
クラス1
未知クラス
ξ_i^{2+}
ξ_i^{1+}
クラス2の
学習データ
$f(x)=w'\phi(x)$
b_2^-
b_2^+
b_1^-
b_1^+
ξ_i^{1-}
ξ_i^{2-}
$f(x)=0$
(原点)
クラス1の
学習データ

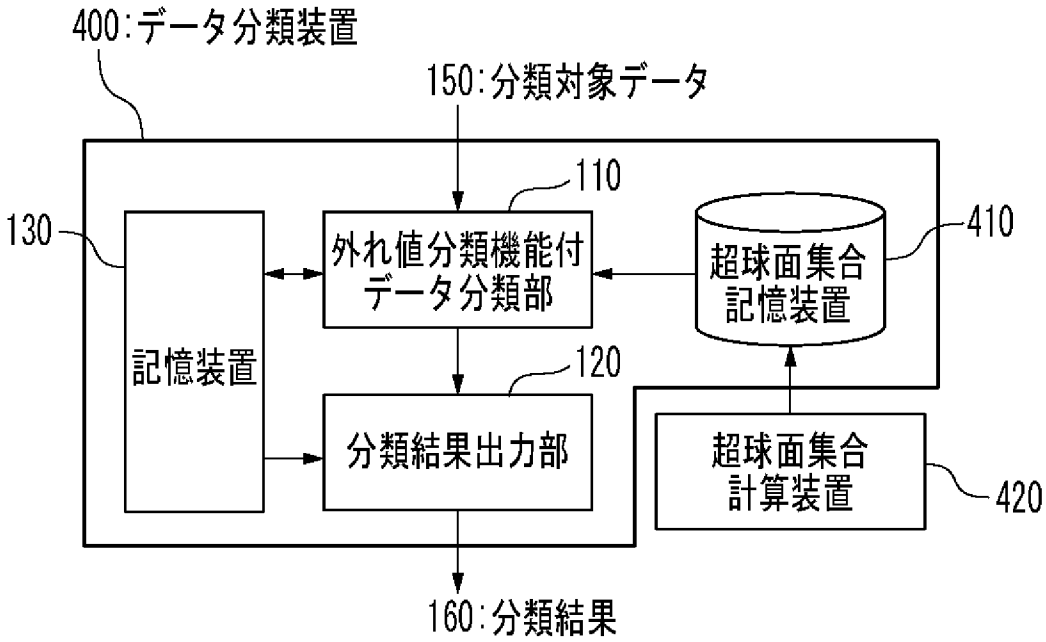
[図13]


未知クラス
クラスl
クラスj
クラスk
b_k^+
b_l^-
b_l^+
b_j^-
b_j^+
b_0^-
b_0^+
b_k^-
f(x)=w' φ(x)
f(x)=0
(原点)

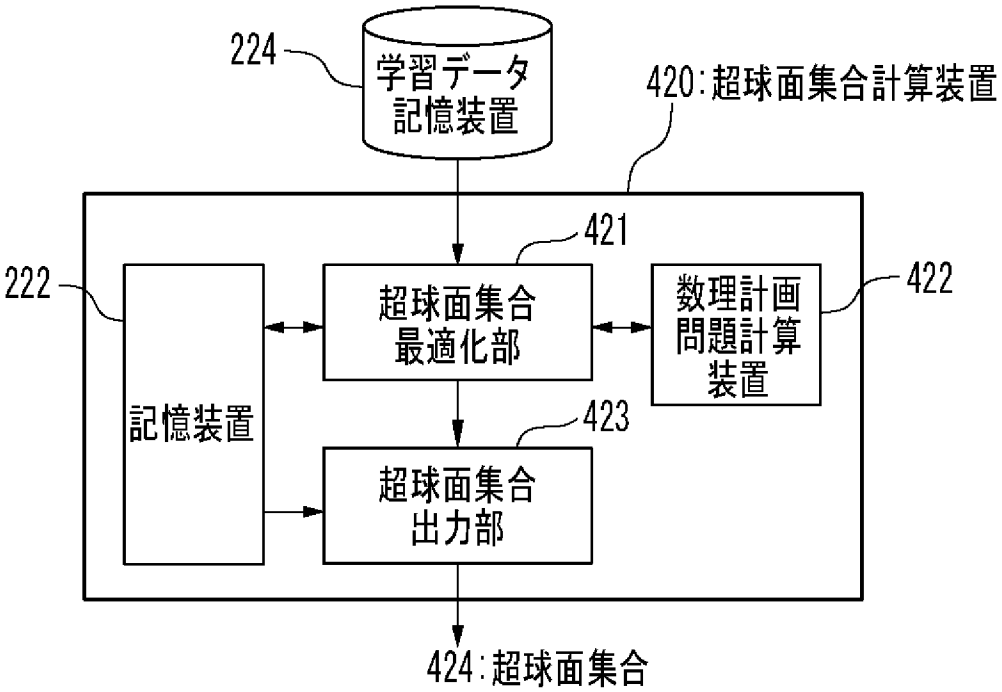
[図14]


この領域で
クラスjとlが重複
クラスl
b_l^-
b_l^+
クラスkが原点を
またいでいる
クラスk
b_k^+
b_0^-
b_0^+
b_k^-
f(x)=w' φ(x)
b_j^-
b_j^+
f(x)=0
クラスj

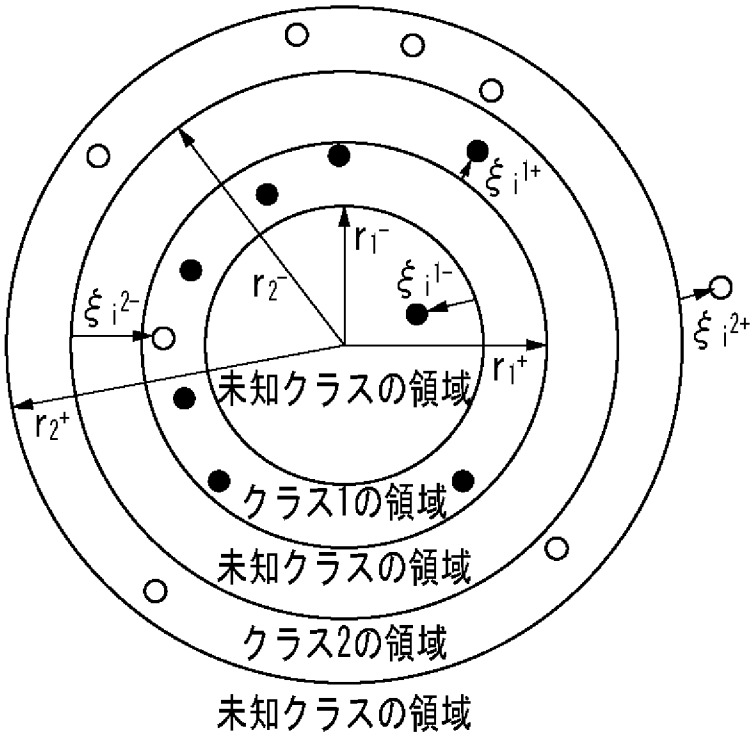
[図15]


400:データ分類装置
150:分類対象データ
110
130
外れ値分類機能付
データ分類部
超球面集合
記憶装置
410
記憶装置
120
分類結果出力部
超球面集合
計算装置
420
160:分類結果

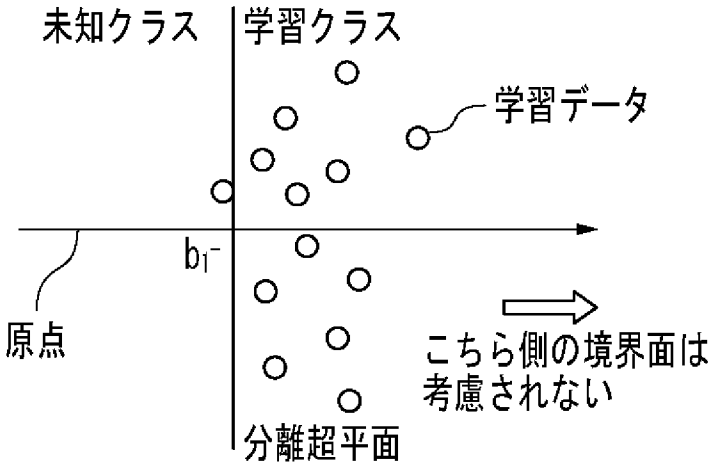
[図16]


224
学習データ
記憶装置
420:超球面集合計算装置
421
222
超球面集合
最適化部
数理計画
問題計算
装置
422
記憶装置
423
超球面集合
出力部
424:超球面集合

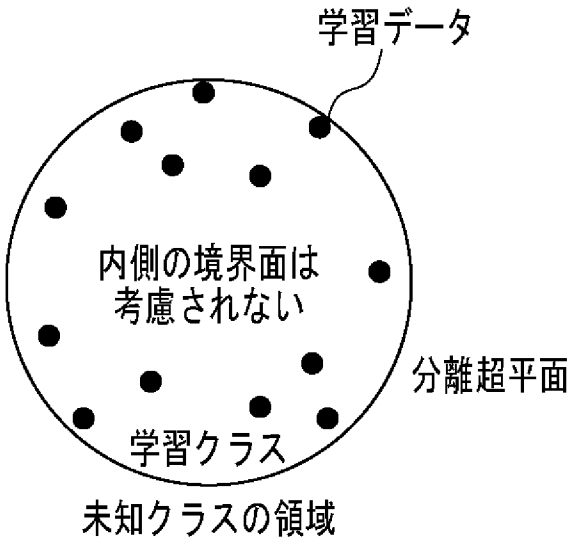
[図17]


ξ_i^{1+}
r_1^-
r_2^-
ξ_i^{1-}
ξ_i^{2-}
ξ_i^{2+}
r_1^+
r_2^+
未知クラスの領域
クラス1の領域
未知クラスの領域
クラス2の領域
未知クラスの領域

[図18]


未知クラス
学習クラス
学習データ
b_1^-
原点
こちら側の境界面は
考慮されない
分離超平面

[図19]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2008/057705

A. CLASSIFICATION OF SUBJECT MATTER
G06F17/30(2006.01)i, *G06N3/00*(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)
G06F17/30, G06N3/00

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.
Y	Yasuto TAKAHATA, "1 Class SVM to Kinbo Support ni yoru Ryoiki Hanbetsu", Keiei no Kagaku Operations Research, Vol.51, No.11, 01 November, 2006 (01.11.06), Vol.51	1-25
Y	JP 2007-52507 A (National Institute of Advanced Industrial Science and Technology), 01 March, 2007 (01.03.07), Par. Nos. [0087] to [0107]; Figs. 8, 15 (Family: none)	1-25

☐ 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	Date of mailing of the international search report
02 May, 2008 (02.05.08)	13 May, 2008 (13.05.08)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２００８／０５７７０５

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F17/30(2006.01)i, G06N3/00(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（ＩＰＣ））

Int.Cl.　G06F17/30, G06N3/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２００８年
日本国実用新案登録公報	１９９６－２００８年
日本国登録実用新案公報	１９９４－２００８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
Y	高畠　泰斗，１クラスＳＶＭと近傍サポートによる領域判別，経営の科学　オペレーションズ・リサーチ　第５１巻　第１１号，2006.11.01，第51巻	1-25
Y	JP 2007-52507 A（独立行政法人産業技術総合研究所）2007.03.01，段落【0087】－【0107】、第8,15図（ファミリーなし）	1-25

☐　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	国際調査報告の発送日
０２．０５．２００８	１３．０５．２００８

国際調査機関の名称及びあて先
日本国特許庁（ＩＳＡ／ＪＰ）
郵便番号１００－８９１５
東京都千代田区霞が関三丁目４番３号

特許庁審査官（権限のある職員）　５Ｂ　３６５０
北川　純次
電話番号　０３－３５８１－１１０１　内線　３５４５