(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6760380号
(P6760380)
(45)発行日　令和2年9月23日(2020.9.23)　(24)登録日　令和2年9月7日(2020.9.7)

| (51)Int.Cl. | | F I | | |
|---|---|---|---|---|
| G01N 30/78 | (2006.01) | G01N | 30/78 | |
| G01N 30/86 | (2006.01) | G01N | 30/86 | G |
| G01N 21/27 | (2006.01) | G01N | 21/27 | Z |
| G06N 99/00 | (2019.01) | G06N | 99/00 | |

請求項の数 9　(全 12 頁)

(21)出願番号　特願2018-531042(P2018-531042)
(86)(22)出願日　平成28年8月3日(2016.8.3)
(86)国際出願番号　PCT/JP2016/072873
(87)国際公開番号　WO2018/025361
(87)国際公開日　平成30年2月8日(2018.2.8)
審査請求日　平成31年1月18日(2019.1.18)

(73)特許権者　000001993
株式会社島津製作所
京都府京都市中京区西ノ京桑原町１番地
(74)代理人　110001069
特許業務法人京都国際特許事務所
(72)発明者　野田　陽
京都府京都市中京区西ノ京桑原町１番地
株式会社島津製作所内

審査官　高田　亜希

最終頁に続く

(54)【発明の名称】分析データ処理方法及び分析データ処理装置

(57)【特許請求の範囲】
【請求項１】
複数の試料の各々について分析装置により収集された、該分析装置が備えるマルチチャンネル検出器の複数のチャンネルの出力値から成る多次元の分析データに対して統計的機械学習を用いた解析手法を適用することにより該分析データを処理する方法であって、
既知の試料について得られた分析データを表す非線形な回帰関数又は判別関数を算出し、
当該算出された非線形な回帰関数又は判別関数の微分値から、該非線形回帰関数又は前記非線形判別関数に対する、前記既知試料の分析データを構成する複数のチャンネルの出力値の各々の寄与度を算出し、
該寄与度に基づき、前記検出器の複数のチャンネルの中から、未知試料について得られた分析データの処理に用いるチャンネルを決定することを特徴とするデータ処理方法。
【請求項２】
請求項１に記載のデータ処理方法において、
前記寄与度に応じて前記既知試料の分析データを構成する複数のチャンネル毎に重み付けを行い、
重み付けを行った後の複数のチャンネルに対して再び寄与度を算出し、該寄与度に基づいて、未知試料について得られた分析データの処理に用いるチャンネルを決定することを特徴とする、データ処理方法。
【請求項３】

請求項１に記載のデータ処理方法において、
決定されたチャンネルに関する情報を提示することを特徴とする、データ処理方法。
【請求項４】
請求項２に記載のデータ処理方法において、
決定されたチャンネルに関する情報を提示することを特徴とする、データ処理方法。
【請求項５】
請求項１に記載のデータ処理方法において、
既知試料について得られた分析データを学習データとテストデータに分け、学習データを用いて、未知試料について得られた分析データの処理に用いるチャンネルを仮決定し、前記仮決定したチャンネルを用いて前記学習データ及び前記テストデータを処理したときの、該学習データ及び該テストデータの適合率の差が所定範囲内にあるときは、前記仮決定したチャンネルを、未知試料について得られた分析データの処理に用いるチャンネルに正式に決定することを特徴とする、データ処理方法。
【請求項６】
請求項２に記載のデータ処理方法において、
既知試料について得られた分析データを学習データとテストデータに分け、学習データを用いて、未知試料について得られた分析データの処理に用いるチャンネルを仮決定し、前記仮決定したチャンネルを用いて前記学習データ及び前記テストデータを処理したときの、該学習データ及び該テストデータの適合率の差が所定範囲内にあるときは、前記仮決定したチャンネルを、未知試料について得られた分析データの処理に用いるチャンネルに正式に決定することを特徴とする、データ処理方法。
【請求項７】
複数の試料の各々について分析装置により収集された、該分析装置が備えるマルチチャンネル検出器の複数のチャンネルの出力値から成る多次元の分析データに対して統計的機械学習を用いた解析手法を適用することにより該分析データを処理する装置であって、
a)既知の試料について得られた分析データを表す非線形な回帰関数又は判別関数を算出する関数算出部と、
b)前記関数算出部で算出された非線形な回帰関数又は判別関数の微分値から、該非線形回帰関数又は前記非線形判別関数に対する、前記既知試料の分析データを構成する複数のチャンネルの出力値の各々の寄与度を算出する寄与度算出部と、
c)前記寄与度に基づき、前記検出器の複数のチャンネルの中から、未知試料について得られた分析データの処理に用いるチャンネルを決定するチャンネル決定部と
を備えることを特徴とするデータ処理装置。
【請求項８】
請求項１に記載のデータ処理方法において、
前記微分値は前記非線形な回帰関数又は判別関数を説明変数で偏微分した値であることを特徴とする、データ処理方法。
【請求項９】
請求項７に記載のデータ処理装置において、
前記微分値は前記非線形な回帰関数又は判別関数を説明変数で偏微分した値であることを特徴とする、データ処理装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、液体クロマトグラフ分析装置やガスクロマトグラフ分析装置、赤外分光光度計、蛍光Ｘ線分析装置等のスペクトル分析装置等、種々の分析装置により収集されたデータを処理する分析データ処理方法及び分析データ処理装置に関する。
【背景技術】
【０００２】
液体クロマトグラフやガスクロマトグラフ等の成分分離装置と検出器とを組み合わせた

クロマトグラフ分析装置では、試料に含まれる多数の成分を時間的に分離した上で該成分を検出器で測定することにより、ある時間（保持時間）における信号強度を示す点データの集合から成る分析データ（クロマトグラムデータ）が得られる。また、検出器として質量分析装置（ＭＳ）を用いたクロマトグラフ質量分析装置（ＬＣ／ＭＳ、ＧＣ／ＭＳ等）では、試料に含まれる成分を時間的に分離した上で、各成分を質量分析装置で測定することにより、ある時間、ある質量電荷比m/zにおける信号強度を示す点データの集合から成る分析データ（クロマトグラムデータ、マススペクトルデータ）を取得することができる。

さらに、赤外分光光度計や蛍光Ｘ線分析装置等のスペクトル分析装置では、試料となる物質に所定の波長範囲の光を照射することにより該物質から放射される光を検出器で測定することにより、ある波長（波数）又はエネルギーにおける信号強度を示す点データの集合から成る分析データ（スペクトルデータ）が得られる。これら分析データを構成する点データの数は、分析装置が備える検出器のチャンネルの数に相当する。

【０００３】

いずれの分析装置においても、分析データから、時間、質量電荷比（m/z）、波長又はエネルギーを横軸とし、検出器のチャンネルの出力（信号強度値）を縦軸とするグラフ（クロマトグラム、マススペクトル、スペクトル）を作成することができる。これらのグラフでは、試料に含まれる成分の種類に応じた位置（保持時間、波長・エネルギー、質量電荷比m/z）にピークが現れる。従って、試料について得られた分析データを解析することにより、該試料の種類や該試料が属するグループなどを識別することができる。

分析装置により収集された分析データから調べたい内容、つまり分析データを解析する目的を変数（目的変数）ｙ、検出器の各チャンネルの出力を変数（説明変数）ｘ１、ｘ２、ｘ３・・・とすると変数ｙは変数ｘ１、ｘ２、ｘ３・・・を使って表すことができる。変数ｘ１、ｘ２、ｘ３・・・は互いに独立した変数であることから、統計学上、上記分析データは変数ｘ１、ｘ２、ｘ３・・・の数だけ次元を有する多次元データとして扱われる。

【０００４】

多種多様な化合物の混合物から成る試料について得られる分析データの場合、グラフには多数のピークが発生するが、全てのピークについてその位置や大きさを解析する作業は効率が悪い。特定のピークに着目することにより、作業の効率化を図ることができるが、どのピークに着目すべきか判断することは困難である。そこで、このような問題を解決する解析手法として、主成分分析（Principal Component Analysis: PCA、非特許文献１）や非負行列因子分解（Nonnegative Matrix Factorization: NMF、非特許文献２）、クラスター分析等の多変量解析が利用されている。

多変量解析では、複数グループの試料について得られた分析データの間で、グラフに現れるピークの位置やピーク形状の比較を行い、その結果に基づき分析データの中から不要な点データを削除したり統合したりすることにより分析データを低次元に写像する。低次元に写像された分析データは、その後、回帰分析や判別分析の手法によりモデル化される。

【０００５】

説明変数の数が少ない二次元データや三次元データ等、比較的単純なデータの場合は線形回帰分析、線形判別分析の手法が適用される。一方、分析装置で得られる分析データのような多次元データの場合は、線形回帰や線形判別の分析手法を適用することが難しく、ニューラルネットやサポートベクターマシン（SVM）等の学習機械による非線形回帰分析、非線形判別分析の手法が適用される。

【０００６】

PCAやNMFでは、分析データをモデル化したときに信号強度が変動する次元を全て反映できるように低次元に写像する。

例えば、蛍光Ｘ線分析装置の検出結果に基づきプラスチックの種類を判別する場合、予め、プラスチックの種類が既知の複数グループについてスペクトルデータを取得し、これ

ら複数グループのデータ間で多変量解析が行われる。スペクトルには、プラスチックのベースとなる材料由来のピーク以外に塗料や可塑剤・難燃剤等の添加物由来のピークが含まれる。一般に、プラスチックの種類が異なると添加物の種類も異なるため、ベース材料由来のピークだけでなく添加物由来のピークも複数のグループのデータ間で変動することになる。従って、この場合はベース材料由来のピークと添加物由来のピークの両方を再現できるように分析データが低次元に写像される。

【０００７】

また例えば、癌疾患の病理マーカを探索するために、健常者グループの生体サンプルについて得られたマススペクトルデータと癌患者グループの生体サンプルについて得られたマススペクトルデータを用いて多変量解析を行う場合、病理マーカとなる成分以外の成分に由来するピークが変動することがある。これは、多くの癌患者に共通する生活習慣（喫煙、飲酒等）があり、該生活習慣に起因する成分に由来するピークも健常者グループと癌患者グループのデータ間で差異があるためである。従って、この場合は癌疾患の病理マーカとなる成分由来のピークと、生活習慣に起因する成分に由来するピークも再現できるように、分析データが低次元に写像される。

【先行技術文献】

【非特許文献】

【０００８】

【非特許文献１】"多変量解析（主成分分析）を活用したクロマトデータ解析",株式会社島津製作所ＨＰ,[平成28年7月25日検索],インターネット<URL:http://www.an.shimadzu.co.jp/hplc/support/lib/lctalk/82/82tec.htm>

【非特許文献２】NGOC-DIEP HO, "NONNEGATIVE MATRIX FACTORIZATIONALGORITHMS AND APPLICATIONS",インターネット<URL:https://www.researchgate.net/profile/Ngoc_Diep_Ho/publication/262258846_Nonnegative_matrix_factorization_algorithms_and_applications/links/02e7e537226cb7e59b000000.pdf>

【非特許文献３】Tomoo AOYAMA and Hiroshi ICHIKAWA, "Obtaining the Correlation Indices between Drug Activity and Structural Parameters Using a Neural Networl", Chem. Pharm. Bull. 39(2) 372-378, (1991)

【非特許文献４】Karen Simonyan et al., "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", インターネット<URL:http://arxiv.org/pdf/1312.6034v2.pdf>

【発明の概要】

【発明が解決しようとする課題】

【０００９】

上述したようにして低次元に写像された後の分析データをニューラルネットやSVM等の学習機械に入力して非線形回帰分析や非線形判別分析の手法を用いてモデル化すると、次のような問題が生じる。

上述した添加物由来のピークや生活習慣に起因する成分由来のピークは、プラスチックの種類や癌疾患の特徴を表すものではなく、ピークの大きさとプラスチックの種類又は癌疾患であるか否か（疾患の状態）の間に因果関係がない。つまり、本来は両者の間に相関はなく、たとえ相関が見られたとしても偽の相関（偽相関）である。そのため、プラスチックの種類や疾患の状態が既知の複数の試料について得られた分析データを、モデル化するための学習データとした場合に、該学習データでは添加物由来のピークとプラスチックの種類、又は生活習慣に起因する成分由来のピークと疾患の状態との間で相関がみられたとしても、解析対象の分析データに同じような相関がみられるとは限らない。その結果、学習データと同じ手法が解析対象の分析データには適合しない、いわゆる過剰適合状態となる。

【００１０】

過剰適合を防ぐためには、偽相関を示す成分由来のピークがランダムノイズと変わらない存在として無視できるほどに多種多様なパターンの分析データを学習データとして非線

形回帰分析、非線形判別分析を行う必要があるが、そのためには膨大な試料を用意する必要があり、現実的ではない。

【００１１】

本発明が解決しようとする課題は、複数の試料について分析装置で収集された多次元データである分析データに基づく、統計的機械学習を用いた解析手法により前記分析データを処理する際に、前記分析データに含まれる、試料の特徴を表す次元を残しつつ、ノイズを排除することである。

【課題を解決するための手段】

【００１２】

一般に、回帰分析又は判別分析に寄与する次元と回帰分析又は判別分析に用いる関数の出力値の相関は高いため、相関が高い次元のみを分析に利用し、中途半端な相関をもつ次元を削除することを考える。当然ながら、全く相関を持たない次元はノイズである可能性が高いため、削除する。

【００１３】

線形回帰分析や線形判別分析では、相関係数を計算で求めることができるが、ニューラルネットやサポートベクターマシン等の機械学習による非線型関数を用いた回帰分析や判別分析では相関係数を求めることができない。ただし、ニューラルネットを用いた回帰・判別分析では、出力に対する入力の各次元の寄与度を偏微分を用いて算出することが可能である（非特許文献４）。なお、非特許文献４には、ニューラルネットによる非線形回帰／判別分析において、シグモイド関数を用いることが記載されているが、シグモイド関数に限ることなくニューラルネットを用いた学習法としては勾配法が一般的であるため、回帰関数又は判別関数の各データ点における偏微分値（又は劣偏微分値）を算出することが可能である。また、サポートベクターマシンを用いた機械学習による回帰／判別分析においても、入力及び出力がともに連続的な値をとる機械学習手法であれば、同様に、偏微分値を算出したり、偏微分に相当する値として、入力を微小に変化させた場合の差分を算出したりすることが可能である。分析データの各データ点における偏微分値又はそれに相当する値を算出することができれば、その値から寄与度を算出することができる。

【００１４】

そこで、本発明は、複数の試料の各々について分析装置により収集された、該分析装置が備えるマルチチャンネル検出器の複数のチャンネルの出力値から成る多次元の分析データに対して統計的機械学習を用いた解析手法を適用することにより該分析データを処理する方法であって、

既知の試料について得られた分析データを表す非線形な回帰関数又は判別関数を算出し、

当該算出された非線形な回帰関数又は判別関数の微分値から、該非線形回帰関数又は前記非線形判別関数に対する、前記既知試料の分析データを構成する複数のチャンネルの出力値の各々の寄与度を算出し、

該寄与度に基づき、前記検出器の複数のチャンネルの中から、未知試料について得られた分析データの処理に用いるチャンネルを決定することを特徴とする。

【００１５】

上記分析データ処理方法において分析装置とは、マルチチャンネル検出器を備えたものであれば何でも良く、代表的なものとして、質量分析装置、液体クロマトグラフ分析装置、ガスクロマトグラフ分析装置、赤外分光光度計、蛍光Ｘ線分析装置等のスペクトル分析装置が挙げられる。

また、既知の試料とは、含まれる成分が既知の試料、プラスチックの種類や癌患者であるか健常者であるか、というように属するグループが既知の試料等をいう。反対に、未知の試料とは、含まれる成分が未知の試料、属するグループが未知の試料をいう。

統計的機械学習には、ニューラルネットやサポートベクターマシン等の学習機械を用いることができる。

既知試料の分析データを表す非線形な回帰関数又は判別関数の微分値は、検出器の各チ

ャンネルの出力値を表す変数（説明変数）で回帰関数又は判別関数を偏微分することにより算出することができるが、算出にかかる時間を低減するために分析データの一部のデータを抜粋したり、分析データをクラスタリングして各クラスターの代表点で代用したり、経験的に求められた標準的なデータパターンに対して微分値を求めたりしても良い。

【００１６】

上記分析データ処理方法において、寄与度に基づきチャンネルを決定する基準は経験的に設定することが可能である。代表的な基準として、例えば寄与度の高い順から上位n個のチャンネルを選択する方法が挙げられる。

この場合、過剰適合が発生しないように、選択するチャンネルの数nを決定すると良い。

過剰適合状態とは、回帰関数又は判別関数を求めるために用いた分析データ自身には、当該回帰関数又は判別関数が適合するが、それ以外の分析データには適合しない状態をいう。例えば、成分が既知の分析データを、回帰関数又は判別関数を求めるための学習データと、学習データについて得られた回帰関数又は判別関数を検証するためのテストデータに分け、学習データについて得られた回帰関数又は判別関数を、学習データ自身に適用した場合の適合率と、前記回帰関数又は判別関数をテストデータに適用した場合の適合率を求め、これらの差が大きければ大きいほど、過剰適合状態にあると判断することができる。

以上より、上記分析データ処理方法においては、既知試料について得られた分析データを学習データとテストデータに分け、学習データを用いて、未知試料について得られた分析データの処理に用いるチャンネルを仮決定し、前記仮決定したチャンネルを用いて前記学習データ及び前記テストデータを処理したときの、該学習データ及び該テストデータの適合率の差が所定範囲内にあるときは、前記仮決定したチャンネルを、未知試料について得られた分析データの処理に用いるチャンネルに正式に決定することが好ましい。

【００１７】

また、上記分析データ処理方法において、好ましくは、検出器の各チャンネルの寄与度に応じて既知試料の分析データを構成する複数のチャンネル毎に重み付けを行い、

重み付けを行った後の複数のチャンネルに対して再び寄与度を算出し、重みを更新することを繰り返す。その重み又は寄与度に基づいて、未知試料について得られた分析データの処理に用いるチャンネルを決定する。

重み付けは、寄与度を強調するような処理、つまり、大きい寄与度はより大きくなるような処理が好ましく、例えば寄与度を累乗する、寄与度の対数をとる、といった処理が挙げられる。また、重みの大きさは、試料の種類や分析装置の種類等に応じて実験的に求めておいても良い。このように重み付けを行う場合も、寄与度から直接チャンネルを決定する場合も、決定されたチャンネルの出力値を用いた機械学習結果に対して再び同様のチャンネル決定を繰り返し行うことにより、チャンネルの数を段階的に減らして行くようにしても良い。

【００１８】

なお、重み付けを行う前の寄与度に基づきチャンネルを決定した場合、重み付けを行った後の寄与度に基づきチャンネルを決定した場合のいずれにおいても、機械学習の結果が機械学習対象となる係数の初期値に依存する場合は寄与度もその初期値による影響を受ける。従って、このような場合は、機械学習を複数回実行した結果に対して得られた複数の寄与度もしくは寄与度に対応した重みの最小値、最大値、平均値を求めてチャンネルの決定に用いてもよい。複数回実行した結果、用いるチャンネルとして決定された回数を基準として正式に用いるチャンネルを決定してもよい。

【００１９】

また、本発明の別の態様は、複数の試料の各々について分析装置により収集された、該分析装置が備えるマルチチャンネル検出器の複数のチャンネルの出力値から成る多次元の分析データに対して統計的機械学習を用いた解析手法を適用することにより該分析データを処理する装置であって、

a)既知の試料について得られた分析データを表す非線形な回帰関数又は判別関数を算出する関数算出部と、
b)前記関数算出部で算出された非線形な回帰関数又は判別関数の微分値から、該非線形回帰関数又は前記非線形判別関数に対する、前記既知試料の分析データを構成する複数のチャンネルの出力値の各々の寄与度を算出する寄与度算出部と、
c)前記寄与度に基づき、前記検出器の複数のチャンネルの中から、未知試料について得られた分析データの処理に用いるチャンネルを決定するチャンネル決定部と
を備えることを特徴とする。
【発明の効果】
【００２０】
本発明に係る分析データ処理方法及び分析データ処理装置によれば、分析データに含まれる複数のチャンネルの出力値のうち、ノイズとなるチャンネルの出力値を排除し、回帰分析・判別分析に寄与するチャンネルの出力値、すなわち試料の特徴を表すチャンネルの出力値を用いて、未知試料の分析データを解析することができる。
【図面の簡単な説明】
【００２１】
【図１】本発明の一実施形態である分析システムの概略構成図。
【図２】データ処理方法の手順を示すフローチャート。
【図３】ＰＰの試料について得られた吸光比スペクトルの一例。
【図４】フルコネクトーニューラルネットワークの概念図。
【図５】分析データの各データ点の寄与度を示す図。
【図６】過剰適合の発生を調べた図。
【図７】重み付けを行った後の寄与度を示す図。
【発明を実施するための形態】
【００２２】
図１は、本発明の一実施形態である分析システムの概略構成図である。
分析システムは、分析装置１０とデータ処理装置２０とから成る。分析装置１０は、計測部１１とマルチチャンネル検出器１２（以下、検出器１２という）と該検出器１２による検出信号をデジタルデータに変換するアナログ-デジタル変換部（ＡＤＣ）１３とを備える。例えば分析装置１０がフーリエ変換赤外分光光度計（ＦＴＩＲ）の場合、計測部１１は、試料に照射する赤外干渉光を生成する干渉計から成り、検出器１２は、ＴＧＳ検出器やＭＣＴ検出器等から成る。
【００２３】
データ処理装置２０は、ＡＤＣ１３においてアナログ-デジタル変換された、検出器１２のチャンネルの出力データに対して所定のデータ処理を行うことで多次元データである分析データを作成するデータ収集部２１と、分析データに基づき赤外吸収スペクトルやクロマトグラム等のグラフを作成するグラフ作成部２２と、前記分析データを解析するデータ解析部２３と、データ解析部２３における解析に用いられる解析用データベース２４と、データ解析部２３において解析された結果を表示する表示部２５と、を備える。
【００２４】
なお、データ処理装置２０の機能は、専用のハードウェアを用いて実現することも可能であるが、汎用のパーソナルコンピュータをハードウェア資源とし、該パーソナルコンピュータにインストールされた専用の処理ソフトウェアを実行することにより実現するのが一般的である。
【００２５】
続いて、上記データ処理装置２０におけるデータ処理方法の手順を図２に示すフローチャートを参照しながら説明する。図２のフローチャートの各ステップの処理はデータ処理装置２０のデータ解析部２３が実行する。なお、以下の説明において「入力データ」とはデータ解析部２３に入力されるデータを指し、「出力データ」とはデータ解析部２３から出力されるデータを指す。

【００２６】
＜ステップ１　入力データの正規化＞
分析装置１０によっては、検出器１２の出力値の再現性が低く、たとえ同一試料であっても、測定する毎に検出器１２の各チャンネルの出力値が異なる場合がある。また、分析装置１０によっては、検出器１２のチャンネル毎に感度やＳＮ比が異なる場合もある。例えば質量分析装置では検出器の再現性が低く、マススペクトルに現れるピークの再現性が低い。また、ＦＴＩＲ等の吸光分析装置では、波長によって検出器の感度やＳＮ比が大きく異なる。
【００２７】
そこで、検出器１２の各チャンネルの出力値の変動量の期待値がほぼ一定になるよう、つまり、検出器１２のチャンネルの出力値の標準偏差が一定になるように正規化する。正規化には種々の周知の方法を用いることができる。例えば、複数の分析データを構成する任意のチャンネルの出力値、つまり、スペクトルやマススペクトル、クロマトグラフ中の任意のピーク値を、その標準偏差で除する処理とすることができる。
【００２８】
＜ステップ２　学習機械を用いた非線形回帰分析又は非線形判別分析＞
学習データについて、ニューラルネットやＳＶＭ等の学習機械を用いた非線形回帰又は非線形判別（学習）を行う。学習データとは、例えば種類が既知の樹脂、癌患者か健常者のいずれであるかが既知の生体サンプルなど、解析結果が既知の試料について分析装置１０から得られた分析データを指す。この場合、解析対象試料の分析データに対して適用する非線形回帰分析又は非線形判別分析と同じ回帰対象変数又は判別ラベルで、学習データの非線形回帰分析又は非線形判別分析を行う。ステップ２の処理により、学習データを表す回帰関数又は判別関数が求められる。
【００２９】
＜ステップ３　入力データの偏微分値の算出＞
学習データについて得られた回帰関数・判別関数を偏微分する。偏微分は、例えば非特許文献４に記載されているような手法を用いることができる。この手法では、softmax関数に入力される値を出力値とみなして微分する。
【００３０】
＜ステップ４　寄与度の算出＞
ステップ３において算出された偏微分値を用いて各チャンネルの寄与度を算出する。例えば、樹脂種を識別するために得られたスペクトルデータのように、特定のチャンネルの信号強度値が大きくなればなるほど、ある物質を含む確度が上がるという場合は、偏微分値は正の値を示すため、偏微分値の平均を取れば良い。一方、例えばある疾病に罹患しているか否かを判断するための病理マーカを調べるためのマススペクトルデータでは、特定のチャンネルの信号値が適正値からどの程度外れているかが重要となる。このような場合は、正負両方の偏微分値が現れるため、偏微分値の二次ノルムから寄与度を算出する。
【００３１】
＜ステップ５　チャンネルの決定＞
ステップ４で算出された寄与度の大きい順にn個のチャンネルを選ぶ。この場合、選択する数nとして一つの値を設定しても良いが、いくつかの値を設定し、既知のテストデータ（学習データとは別の既知の分析データ）について選択したn個のチャンネルの出力値を用いて、ステップ２の回帰・判別分析を行った結果、過剰適合が少なく、チャンネルを減らしたことによる精度低下が少なければ、それらn個のチャンネルを最終的にデータ処理に使用するチャンネルに決定すると良い。
【００３２】
また、上位n個のチャンネルを選択した後、これらn個のチャンネルの出力から成る分析データについて、ステップ２～４の処理を行うと、各チャンネルの寄与度の大きさの順位が入れ替わる場合がある。そこで、まずは、最終的に選択する数nよりも多い数（ｎ＋α）のチャンネルを選択し、それら選択したチャンネルについてステップ２～５の処理を行っ

てチャンネルの数を段階的に減らし、最終的にn個のチャンネルを決定するようにしても良い。これにより、チャンネルの寄与度の順位の入れ替わりの影響を軽減することもできる。

【００３３】

次に本発明を、プラスチック試料についてＦＴＩＲを用いて収集された分析データに基づき、試料の樹脂種の判定のためのデータ処理を行った結果について図３～図７を参照して説明する。

図３、図５及び図６は、添加物など含むＰＰ（ポリプロピレン）、ＰＥ（ポリエチレン）、ＰＵＲ（ポリウレタン樹脂）、ＡＢＳ樹脂（アクリルニトリル－ブタジエン－スチレン共重合合成樹脂）の４種類の樹脂についてＦＴＩＲで得られた分析データ（スペクトルデータ）に基づき、ＰＰと非ＰＰのいずれであるかを識別した結果を示す。

【００３４】

図３は、ＰＰの試料について得られた吸光比スペクトルの一例を示す。このスペクトルは、ＦＴＩＲで得られた吸光比スペクトルを正規化処理（全ての測定点における信号強度値の標準偏差で各測定点の信号強度値を除する処理）したものである。データ解析部２３の関数算出部２３１には、このように正規化した後の吸光比スペクトルデータが入力される。

【００３５】

データ解析部２３では、ＰＰと非ＰＰの樹脂の判別を、図４に示すフルコネクト－ニューラルネットワークを用いて行う。ここでは、中間層の活性化関数としてelu関数を用い、出力層の活性化関数としてsoftmax関数を用いている。その結果、得られた寄与度を図５に示す。図５の横軸及び図３の横軸は、いずれも検出器１２のチャンネルに対応している。

【００３６】

図６が、得られた寄与度に基づき上位n個のチャンネルの出力値から成る学習データとテストデータを用いて、樹脂種を識別したときの正答率（％）を示す。いずれもデータ数は10000である。

ニューラルネットワークの初期値にも依存するが、分析データに含まれる全てのチャンネルの出力値（1000チャンネル）を用いてＰＰか非ＰＰかを識別したときのテストデータの正答率は94.1％、学習データの正答率は99.2％であった。つまり、学習データでは高い正答率が得られたが、テストデータでは、正答率が低下するという過剰適合状態となった。これに対して、チャンネルの数を減らしていくと、学習データの正答率は徐々に低下する一方、テストデータの正答率が上昇する傾向がみられ、チャンネルの数が４０程度で頭打ちになることが分かった。以上より、この実験例では、寄与度の上位４０のチャンネルの出力値を用いることにより過剰適合を抑えて、正答率（識別率）が向上することが分かる。

【００３７】

また、図７は、図５に示す寄与度を４乗する強調処理を加えた上で、正規化する処理（重み付け）を行った結果を示す。図７から分かるように、重み付けを行うことにより、非常に可読性の高い結果が得られる。重み付け処理を行った後の寄与度に基づき、上位４０個のチャンネルの出力値から成る分析データを用いて正答率を求めたところ、テストデータの正答率は95.5％、学習データの正答率は96％であり、過剰適合が抑えられた。このことから、重み付け処理が、過剰適合の防止に有効であることが分かった。

【００３８】

なお、本発明は上記した実施形態に限らず、適宜の変更が可能である。

例えば、正規化する方法としては、ばらつきから求まる変動係数（＝標準偏差／平均値）で除する周知の手法を用いることができる。

【００３９】

上記実施形態では寄与度を４乗する強調処理を加えた上で、平均を１にする正規化をしたが、強調処理は、寄与度の上位n個を選ぶことと類似の概念であり、強調する度合い（

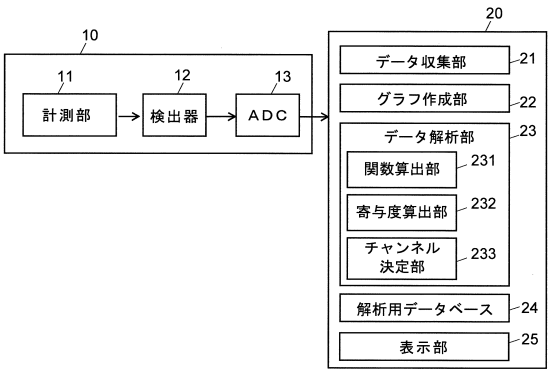
累乗する数）は経験的に調整することが可能である。また、寄与度を累乗することによる強調処理の他、ステップ関数、シグモイド関数などの一般的な非線型な単調関数を用いることができる。

【符号の説明】

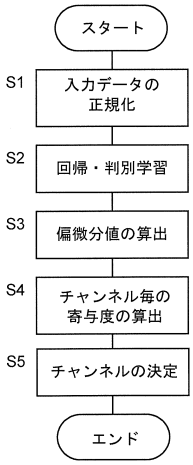
【００４０】

１０…分析装置
１１…計測部
１２…検出器
１３…ＡＤＣ
２０…データ処理装置
２１…データ収集部
２２…グラフ作成部
２３…データ解析部
　２３１…関数算出部
　２３２…寄与度算出部
　２３３…チャンネル決定部
２４…解析用データベース
２５…表示部

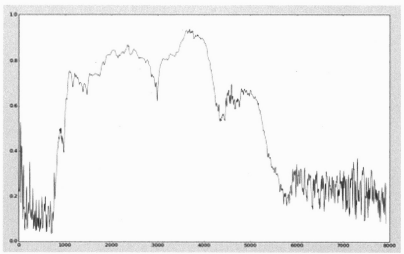
【図１】



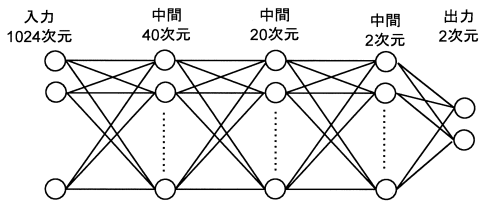

【図２】



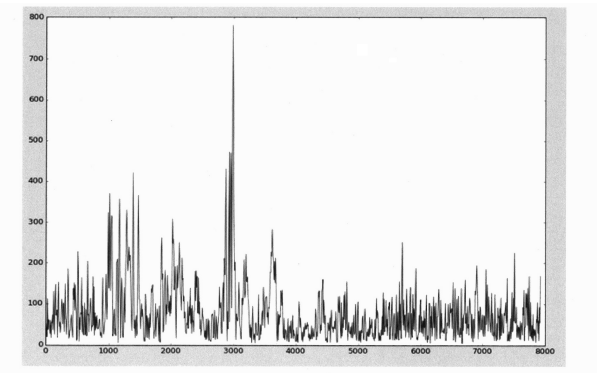

【図３】



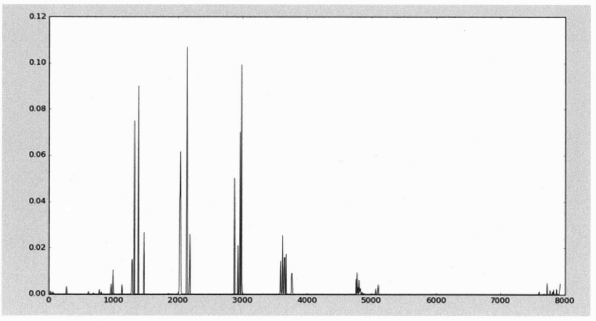
【図４】



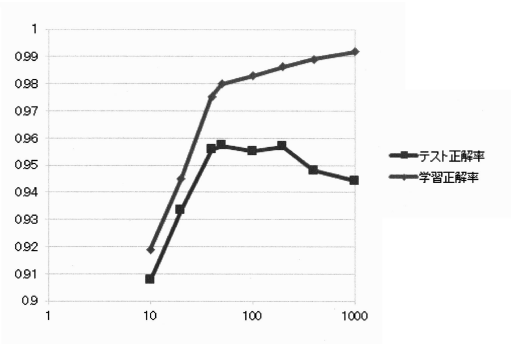
【図５】


800
700
600
500
400
300
200
100
0
0
1000
2000
3000
4000
5000
6000
7000
8000

【図７】


0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
1000
2000
3000
4000
5000
6000
7000
8000

【図６】


1
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
1
10
100
1000
テスト正解率
学習正解率

フロントページの続き

(56)参考文献 米国特許出願公開第２０１３／０２６７７９６（ＵＳ，Ａ１）
特開２０１１－１５０４０８（ＪＰ，Ａ）
Tom Howley et al., The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data, Knowledge-Based Systems, ２００６年, 19, 363-370

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０１Ｎ　３０／００　－３０／９６
Ｂ０１Ｊ　２０／２８１－２０／２９２
Ｇ０１Ｎ　２１／００
Ｇ０６Ｆ　１９／２４
Ｇ０６Ｎ　９９／００
ＪＳＴＰｌｕｓ／ＪＭＥＤＰｌｕｓ／ＪＳＴ７５８０（ＪＤｒｅａｍＩＩＩ）