

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7273844号
(P7273844)

(45)発行日 令和5年5月15日(2023.5.15)

(24)登録日 令和5年5月2日(2023.5.2)

(51)国際特許分類 F I
G 0 1 N 21/27 (2006.01) G 0 1 N 21/27 F

請求項の数 24 (全50頁)

(21)出願番号	特願2020-554164(P2020-554164)	(73)特許権者	516286707
(86)(22)出願日	平成30年4月5日(2018.4.5)		イーエニエーエスセー テック - イン
(65)公表番号	特表2021-526628(P2021-526628 A)		スティチュート デ エンゲンハリア デ
(43)公表日	令和3年10月7日(2021.10.7)		システムス エ コンピュータドレス テ
(86)国際出願番号	PCT/PT2018/050012		クノロジア エ シエンシア
(87)国際公開番号	WO2019/194693		INESC TEC - INSTITU
(87)国際公開日	令和1年10月10日(2019.10.10)		TO DE ENGENHARIA DE
審査請求日	令和3年4月2日(2021.4.2)		SISTEMAS E COMPUTAD
			ORES, TECNOLOGIA E
			CIENCIA
			ポルトガル国 ポルト ルア ドトール ロ
			ベルト フリアス キャンパス ダ ヒューブ
		(74)代理人	100147485
			弁理士 杉村 憲司
		(74)代理人	230118913

最終頁に続く

(54)【発明の名称】 試料からの成分の定量化値を予測する分光測光方法及び装置

(57)【特許請求の範囲】

【請求項1】

定量化される試料からの成分の定量化値を予測する分光測光方法であって、
前記生物試料から電磁スペクトルを取得するステップと、
前記取得したスペクトルを、所定の基底ベクトルによって定義される特徴ベクトルに関連する多次元ベクトル空間である特徴空間内のサンプル点に射影するステップであって、前記次元の各々が予測における特徴であるステップと、
前記特徴空間内のサンプル点から、最小限の個数の近傍のサンプル点が存在すれば当該近傍のサンプル点を選択するステップであって、該近傍のサンプル点には、各々が既知の成分量を有する以前に取得したスペクトルが射影されており、前記最小限の個数は、前記定量化される試料の前記射影したスペクトルと、前記選択した近傍のサンプル点に射影されたスペクトルとの共分散を最大にする数であるステップと、
前記定量化される試料からの前記成分の定量化値を予測するステップであって、前記定量化される試料の前記射影したスペクトルを考慮に入れて、前記選択した近傍のサンプル点からの前記既知の成分量と、前記選択した近傍のサンプル点に射影されたスペクトルとの相関をとることによって、前記定量化値を予測するステップとを含む方法。

【請求項2】

前記定量化される試料の前記成分の定量化値の予測可能性を決定するために、前記成分量の予測誤差の正規分布を、前記選択した近傍のサンプル点から計算し、

前記計算した正規分布から、及び前記定量化される試料の前記射影したスペクトルから p 値を得て、

前記得た p 値を、前記定量化される試料の前記成分の定量化値の予測可能性として用いるステップをさらに含む、請求項 1 に記載の方法。

【請求項 3】

前記最小限の個数の近傍のサンプル点が存在しない場合に、

前記定量化される試料からの前記成分の定量化値の予測が可能でないことをフラグで示すステップをさらに含む、請求項 1 または 2 に記載の方法。

【請求項 4】

前記生物試料から取得したスペクトルを検査データベース内に置くステップと、

前記定量化される試料からの前記成分の、測定した定量化値を受信するステップと、

前記定量化される生物試料から取得した複数のスペクトル、及び前記定量化される試料からの前記成分の、それぞれの前記測定した定量化値を、前記検査データベース内に蓄積するステップと、

前記検査データベース内に蓄積された前記スペクトル及び前記測定した定量化値の部分集合が、前記定量化される試料の前記成分の定量化値の予測を可能にするか否かを判定するステップと、

前記部分集合が予測を可能にする場合に、前記成分の定量化値を予測する前記分光測光方法において用いるために、前記部分集合を前記検査データベースから公開するステップとをさらに含む、請求項 2 または 3 に記載の方法。

【請求項 5】

前記検査データベース内に蓄積された前記スペクトル及び前記測定した定量化値の部分集合が、前記定量化される試料の前記成分の定量化値の予測を可能にするか否かを判定するステップが、

生物試料から取得したスペクトルを前記検査データベース内に置く際に、かつ、

前記生物試料に相当する前記定量化される試料からの前記成分の、前記測定した定量化値を受信した際に、

前記取得したスペクトルを、所定の基底ベクトルによって定義される特徴ベクトルに関連する多次元ベクトル空間である特徴空間内のサンプル点に射影するステップであって、前記次元の各々が予測における特徴であるステップと、

前記検査データベースにおいて、前記特徴空間内のサンプル点から、最小限の個数の近傍のサンプル点が存在すれば当該近傍のサンプル点を選択するステップであって、該近傍のサンプル点には、各々が既知の成分量を有する以前に取得したスペクトルが射影されており、前記最小限の個数は、前記定量化される試料の前記射影したスペクトルと、前記検査データベースにおいて選択した前記近傍のサンプル点に射影されたスペクトルとの共分散を最大にする数であるステップと、

前記定量化される試料からの前記成分の定量化値を予測するステップであって、前記定量化される試料の前記射影したスペクトルを考慮に入れて、前記検査データベースにおいて選択した前記近傍のサンプル点からの前記既知の成分量と、前記検査データベースにおいて選択した前記近傍のサンプル点に射影されたスペクトルとの相関をとることによって、前記定量化値を予測するステップと、

前記定量化される試料の前記成分の定量化値の予測可能性を決定するステップと、

前記予測可能性が所定閾値を上回るものと考えられる場合に、前記成分の定量化値を予測する前記分光測光方法において用いるために、前記選択した近傍のサンプル点を前記検査データベースから公開するステップと

をさらに含む、請求項 4 に記載の方法。

【請求項 6】

前記公開される前記選択した近傍のサンプル点が局所モデルを構成する、請求項 5 に記載の方法。

【請求項 7】

10

20

30

40

50

前記特徴空間内のサンプル点から前記最小限の個数の近傍のサンプル点を選択するステップが、

前記取得したスペクトルを、前記多次元ベクトル空間である特徴空間内のサンプル点に射影するステップと、

複数の探索方向を前記特徴空間内に規定するステップと、

前記特徴空間内に包含される複数の方向性探索体積空間を規定するステップであって、前記方向性探索体積空間の各々は、前記特徴空間内の前記スペクトルを射影したサンプル点を含み、前記スペクトルを射影したサンプル点から1つの探索方向に沿って所定の探索半径の距離だけ広がり、前記探索方向に沿って所定の探索幅の距離だけ広がるステップと、

前記探索方向毎に複数の対応する予測モデルを計算するステップであって、該予測モデルの各々は、次元の部分集合を前記特徴空間の次元から選択することによって計算し、該予測モデルは、前記探索方向に対応する前記方向性探索体積空間内の前記スペクトルを射影したサンプル点を用いて、前記共分散が最大になるように計算するステップと、

前記定量化される成分の定量化値の予測可能性の最大値を有する前記対応する予測モデルを有する前記探索方向を選択するステップと、

前記選択した探索方向に対応する選択した前記方向性探索体積空間内の前記スペクトルを射影したサンプル点を、前記選択した最小限の個数の近傍のサンプル点として用いるステップと

を含む、請求項1～6のいずれかに記載の方法。

【請求項8】

前記方向性探索体積空間の各々が、前記特徴空間内で前記スペクトルを射影したサンプル点から始まる領域として規定される、請求項7に記載の方法。

【請求項9】

前記所定の探索幅の距離を低減することによって前記選択した方向性探索体積空間を最小にするステップであって、前記選択した次元の部分集合によって計算され、かつ前記最小にされる方向性探索体積空間内の前記スペクトルを射影したサンプル点を用いて計算された前記予測モデルによって、前記定量化される成分の定量化値の予測可能性が最大にされるようにするステップをさらに含む、請求項7または8に記載の方法。

【請求項10】

前記特徴空間と前記成分の定量化値との共分散行列を定義し、
 予測誤差を最小にする前記共分散行列から抽出した固有ベクトルの数を最小にし、
 前記最小の数に対応する固有ベクトルを選択し、
 前記選択した固有ベクトルによって定義される多変量線形予測モデルを、前記探索方向の各々について計算した前記予測モデルとして使用することによって、前記探索方向毎の前記予測モデルを計算する、請求項7～9のいずれかに記載の方法。

【請求項11】

前記計算した多変量線形予測モデルが、前記所定の基底ベクトルによって定義される射影と合わせて、入力スペクトルと成分の定量化値との相関を与える、請求項10に記載の方法。

【請求項12】

前記方向性探索体積空間の各々が、それぞれの前記探索方向及び前記スペクトルを射影したサンプル点によって規定される直線からの所定距離によって規定される多次元のボックスまたは円柱である、請求項11に記載の方法。

【請求項13】

前記選択した方向性探索体積空間内の前記スペクトルを射影したサンプル点について、前記選択した方向性探索体積空間に対応する前記予測モデルによって計算した前記定量化される成分の定量化値の変化を直交フィルタリングするステップをさらに含む、請求項7～12のいずれかに記載の方法。

【請求項14】

10

20

30

40

50

前記特徴空間内のサンプル点から前記最小限の個数の近傍のサンプル点を選択することを反復するステップを含み、該反復するステップは、

複数の探索方向を前記特徴空間内に規定するステップと、

前記特徴空間内に包含される複数の方向性探索体積空間を規定するステップであって、前記方向性探索体積空間の各々は、前記特徴空間内で、以前に選択した前記方向性探索体積空間の選択した前記探索方向に沿った前記所定の探索半径の距離の終端から始まる領域として規定されるステップと、

前記探索方向毎に複数の対応する予測モデルを計算するステップであって、該予測モデルの各々は、次元の部分集合を前記特徴空間の次元から選択することによって計算し、該予測モデルは、前記探索方向に対応する前記方向性探索体積空間内の前記スペクトルを射影したサンプル点を用いて計算するステップと、

10

前記定量化される成分の定量化値の予測可能性の最大値を有する前記対応する予測モデルを有する前記探索方向を選択するステップと、

前記選択した探索方向に対応する前記選択した方向性探索体積空間内の前記スペクトルを射影したサンプル点を、前記選択した最小限の個数の近傍のサンプル点として用いるステップと

によって実行する、請求項 7 ~ 13 のいずれかに記載の方法。

【請求項 15】

前記定量化される試料の前記射影したスペクトルと、前記選択した近傍のサンプル点に射影されたスペクトルとの共分散についての所定基準に達するまで、前記反復するステップに含まれるステップを実行するステップを含む、請求項 14 に記載の方法。

20

【請求項 16】

前記特徴空間内のサンプル点から前記最小限の個数の近傍のサンプル点を選択することを反復して、前記予測モデルを再帰的に計算するステップと、

前記計算した予測モデルを統合して、パスモデルを伴う統合予測モデルにするステップとを含む、請求項 14 または 15 に記載の方法。

【請求項 17】

前記パスモデルを、その後に前記予測モデルを計算せずに前記成分の定量化値を予測するためにキャッシュ記憶する、請求項 16 に記載の方法。

【請求項 18】

30

前記予測される定量化値が、前記試料からの成分から決定されるクラスのロジスティック関数である、請求項 7 ~ 17 のいずれかに記載の方法。

【請求項 19】

前記所定の探索半径の距離、前記所定の探索幅の距離、及び前記特徴空間内の前記複数の探索方向の数を、反復最適化方法、特にシプレックス・アルゴリズムを用いて決定する、請求項 7 ~ 18 のいずれかに記載の方法。

【請求項 20】

前記特徴空間内のサンプル点から前記最小限の個数の近傍のサンプル点を選択するステップが、前記共分散が共分散の所定の閾値を上回るように前記最小限の個数のサンプル点を選択するステップを含む、請求項 1 ~ 19 のいずれかに記載の方法。

40

【請求項 21】

前記所定の基底ベクトルが、直交する情報保存性の成分関数への分解または行列因子分解、特に特異値分解 - SVD、ウェーブレット、フーリエ変換、ウェーブレット、またはカーレットである、請求項 1 ~ 20 のいずれかに記載の方法。

【請求項 22】

前記試料が生物試料であり、前記成分が生体の代謝産物であり、特に血中代謝産物である、請求項 1 ~ 21 のいずれかに記載の方法。

【請求項 23】

定量化される試料からの成分の定量化値を予測する分光測光方法を実現するためのプログラム命令を含む非一時的記憶媒体であって、前記プログラム命令は、請求項 1 ~ 22 の

50

いずれかに記載の方法を実行すべく実行可能な命令を含む、非一時的記憶媒体。

【請求項 24】

定量化される試料からの成分の定量化値を予測するための分光測光装置であって、電子データプロセッサ、分光光度計、及びプログラム命令を含む非一時的記憶媒体を具え、前記プログラム命令は、前記電子データプロセッサによって実行されると、請求項 1 ~ 22 のいずれかに記載の分光測光方法を前記電子データプロセッサに実行させる分光測光装置。

【発明の詳細な説明】

【技術分野】

【0001】

技術分野

本発明は、定量化される試料からの成分の定量化値を予測するための分光測光方法及び装置に関するものである。

【背景技術】

【0002】

背景

分光法は、代謝産物（代謝物）の識別にせよ定量化にせよ、代謝産物の間接測定である。各分子または原子は、吸光度または放射、反射率、蛍光性、リン（燐）光、及びラマン散乱によって得られる特性スペクトル指紋（フィンガープリント）を有し；帯域強度は試料濃度に正比例する。

【0003】

純粋な物質または単純な混合物では、スペクトルはわずかな干渉しか有さない。これらの場合には、試料識別は帯域整合（バンド・マッチング）によって直接適用可能であり、強度は濃度に比例する。

【0004】

化学製品または医薬品のようなより複雑な混合物では、スペクトル信号は主要吸光帯と高調波との帯域干渉の結果であり；帯域がオーバーラップ（重複）した連続スペクトルを生じさせる。成分間の干渉の増加は、ピーク強度による定量化の可能性を困難に変える。こうした状況では、代謝産物は好適にはその干渉パターンによって定量化しなければならない（非特許文献 1、非特許文献 2）。

【0005】

さらに、分光技術に応じて、帯域分解能及びスペクトル畳み込み（コンボリューション）が大幅に異なり、従って、スペクトル情報は光学部品の再帰型畳み込み（リカレント・コンボリューション）によって局所的に分布して、低品質の極端な場合には自己相関の強い信号をもたらす。スペクトルは物理的情報及び化学的情報を共に有し、並びに成分間の複雑な干渉パターンを有する。分光法の量子性は、あらゆる純粋化合物についての情報が、強度のいくつかの尺度において、異なる波長に沿って広範囲に及ぶことを意味する。光の波動性により、情報の重ね合わせは、試料成分の帯域間の建設的干渉も相殺的干渉も生じさせる。従って、観測される体液の変化は、局所的な無秩序の変化を伴う強非線形であり、最先端のケモメトリクス（計量化学）、機械学習、あるいは人工知能の方法によって単純にモデル化することはできない。非線形の効果からスペクトル情報を抽出するための理論的サポートを開発することを回避するために、多数のケモメトリクス技術者が、人工ニューラル・ネットワーク（神経回路網）、カーネル法、及びサポートベクターマシンのような機械学習アルゴリズムを適用することを試みた。より複雑なモデル構造が、すべての非線形性を捉えてより良好な予測を提供することができるものと期待された。

【0006】

以前の方法は、分光法のモデル化において次の困難性をもたらす。

【0007】

ビッグデータのスペクトル変動性 - ディープ ANN（deep artificial neural networks：深層人工神経回路網）及び非線形 SVM（support vector machine：サポートベ

10

20

30

40

50

クターマシン)はすべてのデータに適合する複合関数モデルであり、即ち、構造化モノリシック・モデルがデータの集合体向けに生成される。こうしたことは非常に複雑なアーキテクチャをもたらし、信号の局所的に無秩序な性質に起因して、分光法ではビッグデータにとって最良ではない。一旦、新たな局所的变化を有する新たな試料が導入されると、こうしたモデルはスペクトル帯域と組成との正確な共分散を見出すことができない。さらに、大量のデータを供給されなければ、ANN及びSVMはその予測において大幅なバイアスが極度に生じる。これらの方法は、特徴空間がほぼ全体的に表現される際に関心事となったに過ぎず；生物学的変動は極めて広大であるので、厳しい。

【0008】

再学習の計算コスト - ANN及びSVMは大域的方法であるので、一旦、異常値の新たな集合が識別されると、上記の複雑なモデル構造を最適化し直さなければならない。一旦、このことを大きなデータベースで行うと、モデル構造を再計算するために大きな計算リソース(資源)を使用しなければならない(非特許文献3)。

10

【0009】

異常値の検出 - ANN及びSVMの複雑な構造は、新たなスペクトルが異常値である場合に「先験的に」決定することを困難にする。あらゆる新たな結果の予測可能性についての結論を引き出すための明らかな法則が存在しないので、スペクトル測定値が異常値であるか否かを知ることは困難である。このことは、医学、獣医学、さらには物質の正確な定量化が最重要であり、予測の失敗が悲惨な結果を有する危険な工業プロセスでは特に重大である。

20

【0010】

従って、光学系及び分光法の体系的変化を考慮に入れる情報処理技術が、高い計算コストを必要とせずこうした問題を解決しそうである。例えば、大域的スペクトル分散を、変化が体系的である特徴的なグループに分けるための局所的な較正方法が開発された(非特許文献4)。多数の場合に、局所的方法はANN及びSVMよりも優れている(非特許文献5)。局所(重み付き)部分的最小二乗回帰(LW-PLS: locally weighted partial least square)(非特許文献6、7)、LOCAL(非特許文献8)、局所バイアス回帰(非特許文献9)及びCARNA(非特許文献10)、及び局所PLSモデル化法(非特許文献11)のような技法は、複雑度の低減及び安定な較正を提供した。

【0011】

最近の開発の1つは手法「スペクトルベースの学習者(SBL: spectral based learner)」であり、NIR(near infrared: 近赤外)土壌組成のビッグデータをモデル化するために使用された(非特許文献12)。SBLは、最適化主成分(oPC: optimized principal component)を用いて構築された知識ベースに基づき、局所的較正は、oPC次元距離近傍点(例えば、k-近傍法または他の距離測定基準を用いて決定される)を用いることによって得られ(非特許文献13)、oPC次元距離近傍点は、化学組成の類似性によって、組成の二乗平均平方根の差(RSD: root square mean difference)を用いて決定される(非特許文献13)。局所的試料選択は主成分、即ち、より低い周波数または信号ベースラインを有するスペクトル指紋に影響を与える物質の影響のみに基づく。SBLは、より低い濃度を定量化するために常に奮闘し、ここでは情報がスペクトルの小規模な変化中に存在する。

30

40

【先行技術文献】

【非特許文献】

【0012】

【文献】P. Geladi and B. Kowalsky, "Partial least squares regression: a tutorial", *Analytical Chemical Acta*, 185:1-17, 1986
A. Phatak and S. Jong, "The geometry of partial least squares", *Journal of Chemometrics*, 11:311-338, 1997
Huang, G.B. Huan, S. Song, and K. You, "Trends in extreme learning machines: A review", *Neural Networks*, 61:32-48, 2013

50

- L. Ramirez-Lopez, T. Behrensa, K. Schmidt, A. Stevens, J.A.M. Dematte, and T. Scholten, "The spectrum-based learner: A new local approach for modeling soil vis-nir spectra of complex datasets", *Geoderma*, 195-196:268-279, 2013
- D.P. Solomatine, Maskey. M., and Shrestha. D.L., "Instance-based learning compared to other data-driven methods in hydrological forecasting", *Hydrol. Process*, 22:275-287, 2008
- T. Neas, T. Isaksson, and B. Kowalski, "Locally weighted regression and scatter correction for near-infrared reflectance data", *Anal. Chem.*, 62(7):664-673, 1990
- C.D. Christy and S.A. Dyer, "Estimation of soil properties using a combination of spectral and scalar sensor data" 10
- J.S. Shenk, M.O. Westerhaus, and P. Erzaghi, "Local prediction with near infrared multi-product databases", *Journal of Near Infrared Spectroscopy*, 5:223-232, 1997
- T. Fearn and A.M.C. Davies, "Locally-biased regression", *Journal of Near Infrared*, 11(6):467-478, 2003
- A.M.C. Davies and T. Earn, "Quantitative analysis via near infrared databases: compression analysis using restructured nearest infrared and constituent data-deux (carnac-d)", *Journal of Near Infrared*, 14(6):403-441, 2003
- F. Goge, R. Joffre, C. Jolivet, I. Ross, and L. Ranjard, "Optimization criteria in sample selection step of local regression for quantitative analysis of large soil nirs database", *Chemometrics and Intelligent Laboratory Systems*, 110(1):168-176, 2012 20
- L. Ramirez-Lopez, T. Behrens, K. Schmidt, A. Stevens, J.A.M. Dematte, and T. Scholten, "The spectrum-based learner: a new local approach for modelling soil vis-nir spectra of complex datasets", *Geoderma*, 195-196:268-279, 2013
- L. Ramirez-Lopez, T. Behrens, K. Schmidt, R.A. ViscarraRossel, J.A.M. Dematte, and T. Scholten, "Distance and similarity-search metrics for use with soil vis nir spectra", *Geoderma*, 199:43-53, 2013
- U.G. Indahl, "The geometry of pls1 explained property: 10 key noted on mathematical properties of and some alternative algorithmic approaches to pls1 modeling", *Journal of Chemometrics*, 24:168-180, 2014 30
- R.J. Pell, L.S. Ramos, and R. Manne, "The model space in partial least square regression", *Journal of Chemometrics*, 21:165-172, 2007
- S. Wold, M. Hoyc, H. Martens, J. Trygg, F. Westade, J. MacGregor, and B.M. Wise, "The pls model space revisited", *Journal of Chemometrics*, 23:67-68, 2009
- R. Ergon, "Finding y-relevant part of x by use of pcr and pls model reduction methods", *Journal of Chemometrics*, 21:537-546, 2007
- R. Ergon, "Re-interpretation of nipals results solves pls model inconsistency problem", *Journal of Chemometrics*, 23:72-75, 2009 40
- R.C. Martins, V.V. Lopez, P. Valentao, J.C.M.F. Carvalho, P. Isabel, M.T. Amara l, M.T. Batista, P.B. Andrade, and B.M. Silva, "Relevant principal component analysis applied to the characterization of Portuguese heather honey", *Natural Product Research*, 22:1560-1582, 2007

【発明の概要】

【発明が解決しようとする課題】

【0013】

現在最新の方法は、スペクトル定量化の複雑性を技術的に解決することができず、医薬品のような重要な用途において利用されるために必要な正確性及び精度を提供することができない。適正な医療上の決定は、分析グレードデータのみによってサポートされる。本

発明は、上述した技術課題、及び分光法における人工知能及びパターン認識の現在の技術的困難性を克服することを意図し、複雑な変動性及びマルチスケール（多重尺度）の干渉の下でのスペクトル試料の正確な定量化及び分類を提供することを意図した方法及び装置を提示する。

【課題を解決するための手段】

【0014】

本発明は、健康状態における代謝産物分類をスペクトル情報から正確に定量化するための、ビッグデータ自己学習人工知能の方法及び装置に関するものであり、このスペクトル情報には、複雑な生物学的変動性及びマルチスケールのスペクトル干渉が存在する。特に、本発明は、複雑度の高い生物学的スペクトル信号を高次元の特徴空間に分解することを可能にし、高次元の特徴空間内で、各部分空間の局所の特徴と、特定の代謝産物濃度またはカテゴリー条件の両者との正確な相関をとる。こうしたことは、人間の介入を必要としない自己学習法によって実現される。開発した人工知能は、特徴空間変換を実行し、共分散の方向を探索し、局所的な組成スペクトルの相関を最適化することによって新たなデータが供給されると、それ自体の知識ベースを確立することができる。

10

【0015】

これらの方法は、定量化値及び分類の知識マップを共に確立することを可能にし、これらの知識マップは計算性能を高めるためにキャッシュ記憶することができる。特に、直接探索は、代謝組成とスペクトル帯域分散との直接的な線形の対応を可能にするデータ及び次元を特徴空間全体にわたって見出すことから成る。さらに、健康状態の異なるクラスの凸包領域を定義するための同様な方法が体液スペクトルから導出される。こうしたことは、定量化値及び分類の両方についての知識マップの作成をもたらす。

20

【0016】

本発明は、新たな推定値の予測可能性、正確性、及び精度を「先験的に」評価することも可能にする。さらに、本発明は、ビッグデータを用いて大域的特徴空間を再定義する自己学習方法を、高い変動性の下での当該大域的特徴空間の適正な特性化、局所的異常の正確な検出、並びに知識ベースを汚染し得る異常値の正確な検出のために提供する。

【0017】

本発明は、複雑なマルチスケール干渉及び生物学的変動性が存在する、分光分析（X線、UV（ultraviolet：紫外線）、VIS（visible：可視光）、NIR（near infrared：近赤外線）、IR（infrared：赤外線）、far-IR（遠赤外線）、及びマイクロ波）において、あるいは他の何らかの種類の分光法（吸光度、反射率、蛍光性、リン光性、ラマン散乱）と共に使用される電磁スペクトルの全領域に適用可能である。本発明は、ヘルスケア（健康管理）、獣医学、バイオテクノロジー（生命工学）、医薬、食品、及び農業のような分野における、非破壊的、非侵襲的な分光法用途の分野にさらに拡張される。

30

【0018】

定量化される試料からの成分の定量化値を予測する分光測光方法を開示し、この方法は：
この生物（生体）試料から電磁スペクトルを取得するステップと；

この取得したスペクトルを、所定の基底ベクトルによって定義される特徴ベクトルに関連する多次元ベクトル空間、ここでは特徴空間内のサンプル点に射影するステップであって、これらの次元の各々が予測における特徴であるステップと；

40

上記特徴空間内のサンプル点から、最小限の個数の近傍のサンプル点が存在すれば当該近傍のサンプル点を選択するステップであって、これらのサンプル点は、各々が既知の成分量を有する以前に取得したスペクトルから射影されており、上記最小限の個数は、上記定量化される試料の上記射影したスペクトルと、上記選択した近傍のサンプル点に射影されたスペクトルとの共分散を最大にする数であるステップと；

定量化される試料からの成分の定量化値を予測するステップであって、定量化される試料の上記射影したスペクトルを考慮に入れて、上記選択した近傍のサンプル点からの既知の成分量と、上記選択した近傍のサンプル点に射影されたスペクトルとの相関をとることによって上記定量化値を予測するステップとを含む。

50

【 0 0 1 9 】

1つの好適例は、定量化される試料の成分の定量化値の予測可能性を決定するために：
成分量の予測誤差の正規分布を、上記選択した近傍のサンプル点から計算し；
上記計算した正規分布から、及び定量化される試料の上記射影したスペクトルから p 値を得て；

得た p 値を、定量化される試料の成分の定量化値の予測可能性として用いることを含む。

【 0 0 2 0 】

1つの好適例は、最小数個の近傍のサンプル点が存在しない場合に：
定量化される試料からの成分の定量化値の予測が可能でないことをフラグで示すステップを含む。

10

【 0 0 2 1 】

1つの好適例は：
上記生物試料から取得したスペクトルを検査データベース内に置くステップと；
定量化される試料からの成分の、測定した定量化値を受信するステップと；
定量化される生物試料から取得した複数のスペクトル、及び定量化される試料からの成分の、それぞれの測定した定量化値を、上記検査データベース内に蓄積するステップと；
上記検査データベース内に蓄積されたスペクトル及び定量化値の測定値の部分集合が、定量化される試料の成分の定量化値の予測を可能にするか否かを判定するステップと；
この部分集合が予測を可能にする場合に、成分の定量化値を予測する本発明の分光測光方法において用いるために、この部分集合を検査データベースから公開するステップとを含む。

20

【 0 0 2 2 】

1つの好適例は、上記蓄積されたスペクトル及び測定した定量化値の部分集合が、定量化される試料の成分の定量化値の予測を可能にするか否かを判定するために：
生物試料から取得したスペクトルを検査データベース内に置く際に；かつ、
上記生物試料に相当する上記定量化される試料からの成分の、上記測定した定量化値を受信した際に；

上記取得したスペクトルを、所定の基底ベクトルによって定義される特徴ベクトルに関連する多次元ベクトル空間、ここでは特徴空間内のサンプル点に射影するステップであって、これらの次元の各々が予測における特徴であるステップと；

30

上記検査データベースにおいて、上記特徴空間内のサンプル点から、最小限の個数の近傍のサンプル点が存在すれば当該近傍のサンプル点を選択するステップであって、これらのサンプル点には、各々が既知の成分量を有する以前に取得したスペクトルが射影されており、上記最小限の個数は、上記定量化される試料の上記射影したスペクトルと、上記検査データベースにおいて選択した近傍のサンプル点に射影されたスペクトルとの共分散を最大にする数であるステップと；

定量化される試料からの成分の定量化値を予測するステップであって、定量化される試料の上記射影したスペクトルを考慮に入れて、上記検査データベースにおいて選択した上記近傍のサンプル点からの既知の成分量と、上記検査データベースにおいて選択した上記近傍のサンプル点に射影されたスペクトルとの相関をとることによって、定量化値を予測するステップと；

40

定量化される試料の成分の定量化値の予測可能性を決定するステップと；

この予測可能性が所定閾値を上回るものと考えられる場合に、成分の定量化値を予測する本発明の分光測光方法において用いるために、上記選択した近傍のサンプル点を上記検査データベースから公開するステップとを含む。

【 0 0 2 3 】

1つの好適例では、上記公開される選択した近傍のサンプル点が局所モデルを構成する。

【 0 0 2 4 】

1つの好適例では、上記特徴空間内のサンプル点から前記最小限の個数の近傍のサンプル点を選択するステップが；

50

取得したスペクトルを、多次元ベクトル空間、ここでは特徴空間内のサンプル点に射影するステップと；

複数の探索方向を上記特徴空間内に規定するステップと；

上記特徴空間内に包含される複数の方向性探索体積空間を規定するステップであって、各方向性探索体積空間は、上記特徴空間内の上記スペクトルを射影したサンプル点を含み、このスペクトルを射影したサンプル点から1つの探索方向に沿って所定の探索半径の距離だけ広がり、この探索方向に沿って所定の探索幅の距離だけ広がるステップと；

探索方向毎に複数の対応する予測モデルを計算するステップであって、各予測モデルは、次元の部分集合を上記特徴空間の次元から選択することによって計算し、予測モデルは、探索方向に対応する上記方向性探索体積空間内の上記スペクトルを射影したサンプル点を用いて、共分散が最大になるように計算するステップと；

10

定量化される成分の定量化値の予測可能性の最大値を有する対応する予測モデルを有する探索方向を選択するステップと；

上記選択した探索方向に対応する選択した方向性探索体積空間内の上記スペクトルを射影したサンプル点を、上記選択した最小限の個数の近傍のサンプル点として用いるステップとを含む。

【0025】

1つの好適例では、各方向性探索体積空間が、上記特徴空間内で上記スペクトルを射影したサンプル点から始まる領域として規定される。

【0026】

1つの好適例は：

上記所定の探索幅の距離を低減することによって上記選択した方向性探索体積空間を最小にするステップであって、上記選択した次元の部分集合によって計算され、かつ上記最小にされる方向性探索体積空間内の上記スペクトルを射影したサンプル点を用いて計算されたモデルによって、定量化される成分の定量化値の予測可能性が最大にされるようにするステップを含む。

20

【0027】

1つの好適例では、上記探索方向毎の予測モデルを：

上記特徴空間と上記成分の定量化値との共分散行列を定義し；

予測誤差を最小にする共分散行列から抽出した固有ベクトルの数を最小にし；

この最小の数に対応する固有ベクトルを選択し；

選択した固有ベクトルによって定義される多変量線形予測モデルを、各探索方向について計算した予測モデルとして使用する

30

ことによって計算する。

【0028】

1つの好適例では、上記計算した多変量線形予測モデルが、上記所定の基底ベクトルによって定義される射影と合わせて、入力スペクトルと成分の定量化値との解釈的な相関を与える。

【0029】

1つの好適例では、上記方向性探索体積空間の各々が、それぞれの探索方向及びスペクトルを射影したサンプル点によって規定される直線からの所定距離によって規定される多次元のボックス（箱形）または円柱である。

40

【0030】

1つの好適例は：

上記選択した方向性探索体積空間内の上記スペクトルを射影したサンプル点について、上記選択した方向性探索体積空間に対応するモデルによって計算した定量化される成分の定量化値の変化を直交フィルタリング（フィルタ処理）するステップを含む。

【0031】

1つの好適例は、上記特徴空間内のサンプル点から最小限の個数の近傍のサンプル点を選択することを反復するステップを含み、この反復するステップは：

50

複数の探索方向を上記特徴空間内に規定するステップと；

上記特徴空間内に包含される複数の方向性探索体積空間を規定するステップであって、各方向性探索体積空間は、上記特徴空間内で、以前に選択した方向性探索体積空間の選択した探索方向に沿った上記探索半径の距離の終端から始まる領域として規定されるステップと；

探索方向毎に複数の対応する予測モデルを計算するステップであって、各予測モデルは、次元の部分集合を上記特徴空間の次元から選択することによって計算し、予測モデルは、探索方向に対応する上記方向性探索体積空間内の上記スペクトルを射影したサンプル点を用いて計算するステップと；

定量化される成分の定量化値の予測可能性の最大値を有する対応する予測モデルを有する探索方向を選択するステップと；

10

上記選択した探索方向に対応する選択した方向性探索体積空間内の上記スペクトルを射影したサンプル点を、上記選択した最小限の個数の近傍のサンプル点として用いるステップと

によって実行する。

【0032】

1つの好適例は、上記のステップを、定量化される試料の上記射影したスペクトルと、上記選択した近傍のサンプル点に射影されたスペクトルとの共分散についての所定基準に達するまで実行するステップを含む。

【0033】

20

1つの好適例は：

上記特徴空間内のサンプル点から、最小限の個数の近傍のサンプル点を選択することを反復して、上記予測モデルを再帰的に計算するステップと；

上記計算した予測モデルを統合して、パス（経路）モデルを伴う統合予測モデルにするステップとを含む。

【0034】

1つの好適例では、上記パスモデルを、その後に予測モデルを計算せずに成分の定量化値を予測するためにキャッシュ記憶する。

【0035】

30

1つの好適例では、予測される定量化値が、上記試料からの成分から決定されるクラスのロジスティック関数である。

【0036】

1つの好適例では、上記所定の探索半径の距離、上記所定の探索幅の距離、及び上記特徴空間内の上記複数の探索方向の数を、反復最適化方法、特にシンプレックス・アルゴリズムを用いて決定する。

【0037】

1つの好適例では、上記特徴空間内のサンプル点から最小限の個数の近傍のサンプル点を選択するステップが、共分散の所定の閾値を上回るように最小限の個数のサンプル点を選択するステップを含む。

40

【0038】

1つの好適例では、上記所定の基底ベクトルが、直交する情報保存性の成分関数への分解または行列因子分解、特に特異値分解 - SVD (singular value decomposition)、ウェーブレット、フーリエ変換、ウェーブレット、またはカーブレット (curvelet) である。

【0039】

1つの好適例では、上記取得したスペクトルの前処理が、これらのスペクトルのデコンボリューション (deconvolution) 及び/または分解能向上を含む。

【0040】

1つの好適例では、上記試料が生物試料であり、上記成分が生体の代謝産物であり、特

50

に血中代謝産物である。

【0041】

非一時的記憶媒体も開示し、この非一時的記憶媒体は、定量化される試料からの成分の定量化値を予測する分光測光方法を実現するためのプログラム命令を含み、これらのプログラム命令は、開示する好適例のいずれかの方法を実行すべく実行可能な命令を含む。

【0042】

定量化される試料からの成分の定量化値を予測するための分光測光装置も開示し、この装置は、開示する好適例のいずれかの分光測光方法を実行するように構成された電子データプロセッサを具えている。

【0043】

好適例は、分光光度計、及び上記分光測光方法を実現するためのプログラム命令を含む非一時的記憶媒体を具えている。

【図面の簡単な説明】

【0044】

【図1】図1 aはスペクトルの集合を示し、図1 bは、ユークリッド距離がスペクトルの特徴の良好な尺度でなく、組成との相関をとることもできない理由を示し、図1 c、dは、最適な射影を可能にする近傍点の集合を特徴空間全体にわたって探索することを示す。

【図2】図2 aは、従来技術（スペクトルに基づく学習者）において用いられたクラスタリング技術を用いる際に、バイアスを有し異常値を有し得る準最適な射影を常に生じさせることを示し、図2 bは、未知のスペクトルにとって最適な射影を示す。

【図3】特徴空間における特徴空間マップに関する問題を示す図である。

【図4】ステップ的な順序プロトコルを示す図である。

【図5】多変量のノ複合的な診断を示す図である。

【図6】局所的なデータ集合の複雑度を低減する方法を示す図である。

【図7】直交フィルタリングが、複雑度のより低い固有構造を生じさせる様子を示す図である。

【図8】自己学習プロセスの主な機構を示す図である。

【図9】特徴空間変換を実行する方法を示す図である。

【図10】特徴空間変換のフルクソグラムを示す図である。

【図11】図11 a、bは、キャッシュ記憶されているモデルを用いて予測を加速する方法を示す図である。

【図12】キャッシュ記憶されているモデルを用いて予測を加速する方法を示す図である。

【図13】PLSが血液のような生体液の複雑性に対処することができない理由を例示する図である。

【図14】血清中のビリルビン及びミオグロビンについての結果を示す図である。

【図15】PLSのベンチマーク対本発明の自己学習人工知能を提示する図である。

【図16】提案する自己学習方法の有効性を各症状について実証する図である。

【図17】特徴空間から最適な局所モデル及び知識ベースを作成する手順を示す図である。

【発明を実施するための形態】

【0045】

詳細な説明

射影モデル

(外1)

$Y = f(X)$

を用いて知識ベース

(外2)

X

10

20

30

40

50

(外2、以下「 X 」とする)
から正確な定量化値
(外3)

Y

(外3、以下「 Y 」とする)
を得ることは：i)共分散
(外4)

$(X^t Y)$

10

(外4、以下「 $(X^t Y)$ 」とする)
が安定であり；ii)スペクトル特徴空間の分散
(外5)

$(X^t X)$

(外5、以下「 $(X^t X)$ 」とする)
が安定であり；iii)予測した
(外6)

20

\hat{Y}

(外6、以下「 \hat{Y} 」とする)
のバイアス分散が低く；iv)抽出した固有ベクトル、射影、及び係数が統計的にコヒーレント（可干渉性）であり解釈可能である場合に実現可能になる。

【0046】

ビッグデータの生体スペクトルの特徴空間全体にわたって大域的に安定な
(外7)

30

$X^t Y$

(外7、以下「 $X^t Y$ 」とする)
及び
(外8)

$X^t X$

(外8、以下「 $X^t X$ 」とする)
は存在しない。共分散の方向は非線形であり、固有ベクトルは局所的特性に応じて特徴空間全体にわたって回転する。 X のとり得る観測値が無数個であるものとするれば、特徴空間の分散が特徴空間全体にわたって非線形である場合、新たな未知のスペクトル
(外9)

40

X_{new}

(外9、以下「 X_{new} 」とする)
を定量化することは実現不可能になる。

【0047】

50

こうした物理的制約があるものとして、本開示中にはスペクトル予測に関する方法を開示し、この方法は、あらゆる未知のスペクトル X_{new} が、局所的部分空間において特徴空間と整合性があるはずであり、このため、組成データとの共分散 ($X^t Y$) に関して整合性のある情報を保有することもできる、ということに基づく。 \hat{Y} の予測が、今度は、 Y についての対応する情報を保有する $X^t X$ の整合性のある部分空間を見出すという問題となり、このため、 $X^t Y$ は X_{new} の変化と一致して、安定した信頼性のある予測を生成する。さらに、 $X^t X$ 及び $X^t Y$ が局所的にコヒーレントであることを保証するだけで、あらゆる未知のスペクトル X_{new} を以前の知識に基づいて予測することができるか否かを「先験的に」知ることが可能になる。

【0048】

与えられた未知の X_{new} に対して物質を定量化するための「先験的」モデルは存在しないものと考えることができる、というのは、「先験的」モデルは以前の方法 (PCR (principal component regression: 主成分回帰)、PLS、LS-SVM、ANN、及び深層学習 (ディープ・ラーニング)) で発生するからである。あらゆる所定の X_{new} について、 \hat{Y} を予測するための整合性を維持することができる知識ベースの特徴空間の部分集合が存在するものと仮定することができる。従って、一旦、新たなスペクトルが記録されると、AI (artificial intelligence: 人工知能) は、改良された基準により正確な予測を実行することを可能にする部分空間が知識ベース内に存在するか否かを学習しなければならない。

【0049】

分光法において部分空間識別を用いることには、次の大きな利点が存在する: i) 複雑度の低減により部分空間の解釈が実現可能になる; データ表現性 (データの数) の局所的独立、即ち、知識ベース内のより多数のデータによって予測が影響されなくなる; iii) 局所的なマルチスケールの整合性; iv) 定量化を実行するために使用する帯域の解釈; iv) 定量化においてどの帯域を用いるかのより良好な制御; v) スペクトル補正がより正確になる、というのは、スペクトル分散に整合性がある場合に、ベースライン散乱、ミー (Mie) 散乱、及びレイリー (Rayleigh) 散乱の補正がスペクトル帯域変動を高めるからである。特徴空間変換 (例えば、カーネル、導関数、ウェーブレット) が局所的に整合性を有する; 及び vi) 適応性: 定量化値が自己学習されるので、局所的適応が、スペクトル

(外10)

(X)

の最適な集合を常に知識ベース内に見出し、この最適な集合は、 \hat{Y} の最良の予測を提供する。部分空間識別は、AI がモデル構築中に自己学習して人間の指示から独立することを可能にする。

【0050】

複素スペクトル情報に基づく定量化の問題を経験的に図1中に説明する。図1aはスペクトルの集合を示し、この組成は異なる量の同じ成分の混合物である。この混合物中に存在する物質は互いに高度に干渉して、濃度とのピーク相関を直接導出して単純な定量化方法を提供することを不可能にする。それにもかかわらず、非線形のスペクトル変化の4つのクラスが存在し、即ち、特定物質の組成との直接の相関を与えることができる4つのモードが存在する。こうした単純な例の下でも、GLM (generalized linear model: 一般化線形モデル) (例えば PLS) の使用は \hat{Y} の予測を高い分散で提供する。さらに、変化のいずれかのモードが表現性に欠ける場合に、このクラス内の新たなスペクトルの予測を提供することは、バイアスの高い予測を提供することが不可避である。図1bは、ユークリッド距離がスペクトルの特徴の良好な尺度でなく、組成との相関をとることもできない理由を示す。図1aの変化の4グループの全部が完全に異なる非線形の射影を呈し、これらの射影は定量化を可能にするが、その定量化値は濃度との線形相関をとれないことがある。複雑な混合物における分光法による定量化は、予測の最小のバイアス分散を局所

10

20

30

40

50

的に生成する固有ベクトルの共変動射影の非線形探索である。

【 0 0 5 1 】

従って、与えられた X_{new} の正確な定量化値を提供するためには、図 1 d 中に提示するように、最適な射影を可能にする近傍点の集合を特徴空間全体にわたって探索しなければならない。

【 0 0 5 2 】

図 2 a は、従来技術（スペクトルに基づく学習者）において用いられた階層的クラスタリング（クラスタ分析）、 k -近傍法（ KNN : k -nearest neighbor）のようなクラスタリング技術を用いる際に、バイアスを有し異常値を有し得る準最適な射影を常に生じさせることを示す。図 2 b は、未知のスペクトル番号 1 及び番号 3 にとって最適な射影を示し、これらのスペクトルは異なる局所的共変動の下であり、番号 2 は異常値であり、知識ベースで予測することはできない。

10

【 0 0 5 3 】

本明細書中には、分光法のビッグデータ用の自己学習方法及び装置を開示する。新たな方法は、整合性のある $X^t X$ 及び $X^t Y$ を維持する定量化値のコヒーレントな固有ベクトルを見出すことができる。提案する自己学習は、モノリシック・モデルを生成しない。新たなデータ毎に、システムは、 X_{new} を射影して \hat{Y} を推定するために、 $X^t X$ 及び $X^t Y$ のコヒーレンスを学習しなければならない。さらに、 $X^t X$ 及び $X^t Y$ が共に局所的にコヒーレントであれば、あらゆる X_{new} の予測問題を「先験的に」推定することができる。分散 - 共分散の整合性についての測定基準は、予測可能性における局所的信頼性を推測することを可能にする。

20

【 0 0 5 4 】

開示する方法の実施形態は、次の 3 つの主要なステップを含む：i) 局所的形状及び部分空間の識別 - スペクトル情報の局所的形状は特徴的部分空間として抽出され、この特徴的部分空間は、局所的な定量化 / 分類をサポートする特徴的な固有ベクトルを有する；ii) 非線形の特徴マッピングの知識ベースを構築する - 局所的形状及び部分空間の反復的識別を適用することが、人工知能の知識ベースをスペクトル情報の非線形マッピング上に構築することを可能にするプロセス；iii) スペクトル情報の局所的最適化 - Y 及び X の両者中、あるいはそれらに対応する特徴空間変換

(外 1 1)

30

K

(外 1 1、以下「K」とする)

及び

(外 1 2)

F

(外 1 2、以下「F」とする)

の両者中の無関係な情報をフィルタリングで除外することにより、凸包体積空間及び予測誤差を最小化することによって、定量化または分類を局所的に改良するプロセス。

40

【 0 0 5 5 】

以下のことは局所的形状及び部分空間の識別に関連する。前に提示したように、定量化値のコヒーレントな固有値を維持することができるデータの局所的な方向クラスタリングが常に存在するものと仮定とする。こうした局所クラスタは、広大な非線形特徴空間内の局所的なモードの変化を表す。従って、 n 次元の特徴空間 F を考え、この特徴空間の座標はスペクトルの特徴の線形結合に比例し、スペクトルの特徴は試料の組成と暗示的に相関がある。この特徴空間の局所的な点における離散的な有限個の方向は、スペクトルの変化のモードを表すことも仮定し、これらのスペクトルは標本の濃度の局所的レベルとコヒーレントである。こうしたことは、スペクトル X と組成 Y との局所的共分散 ($X^t Y$) からコヒー

50

レントな固有ベクトルを抽出することを可能にする。さらに、高度に非線形な特徴空間は、超次元の多面体によって局所的に特性化することができ、この多面体は整合性のある定量化値の方向を有し、即ち、対応する凸包によって包含されるデータの全部が1つのモードを示し、このモードでは、上記特徴空間内部の全スペクトルが、異なるパラメータを低バイアスの分散で定量化する変化のモードに従う。

【0056】

従って、局所的形状は、この局所的部分空間にある組成に直接関係し、定量化値の最適な固有ベクトルを見出すことができる。ビッグデータのスペクトルの定量化及び分類の問題は：i)上記多面体の方向/次元の数を最小にし；ii)バイアス分散を最小にする主方向を獲得し；iii)選択した最適な方向の多面体の凸包の体積空間を最小にする局所的形状の探索に帰し、このため、直接的な線形モデルが、この有限空間近似に適用可能である。

10

【0057】

図3は、特徴空間における特徴空間マップに関する問題を示す。本例では、2次元の特徴空間を提示し、この特徴空間は、空間の異なる領域を占める異なるクラスの高度に非線形な変化を有する。連続線は、標本濃度とのコヒーレントな共分散の特徴を表し、即ち、この線に沿って、局所的な $X^t Y$ のコヒーレントな固有値を見出して、 \hat{Y} の低誤差の推定値を生成することができる。さらに、この線は、この特徴空間の自己学習の特性を表す。例えば、この線の付近に射影されるすべての新たな X_{new} は、自己学習部分空間モデルによって直接予測することができる。

【0058】

この自己学習プロセスは、図3に示すように、低いバイアス及び分散での標本の定量化を可能にするコヒーレントな多面体の部分空間を探索することに焦点を当て、このプロセスの疑似コードはアルゴリズム1に提示する(図17aも参照)。新たな

20

x_i

(外13、以下「 x_i 」とする)

がこの特徴空間内に射影されるものと仮定する。一旦射影されると、この自己学習プロセスは、 Y に関係する x_i の変化の方向内にある最寄りの近傍点を見出さなければならない。従って、この特徴空間内で定量化にとって適正な方向を与える凸包を探索する必要がある。こうした凸包は、 \hat{Y} を予測する局所的な $X^t Y$ の最小の体積空間及び最小数の固有ベクトルも提示しなければならない。以下の手順は、この自己学習プロセスの具体例を記述する：

30

ステップA．方向を見出す

目的： x_i の最小数の方向及び局所的部分空間形状を見出す。

1．初期化：i)ある探索半径を有する円形領域を x_i の射影の周りに規定する；ii)方向の数を規定する；iii)各方向の探索の次元を規定する。

2．初期の探索：i)最適な固有ベクトルの数を決定して局所モデルの誤差を予測する；ii)統計的に不整合な方向を除外する；

40

3．新たな反復を用意する：i)整合性のある方向内で、最悪の寄与を除外し、探索の長さを特定の方法で削除するか増加させる；ii)各凸包の極点及び x_i を考慮に入れて、新たな探索方向を計算する；

4．探索ループ：i)新たな方向の固有ベクトル数及び予測誤差を特定する；ii)最悪の方向を除外する；iii)最悪の(より小さい、またはより大きい)の長さを除外することによって探索の次元を再設定し、それに応じて探索の長さを増加または減少させる；iii)統計的に有意な方向または次元の変化が発生しなくなるまで、前の動作を繰り返す。

5．出力：最小数の実現可能な方向、及び各方向の凸包体積空間。

ステップB．凸包の最適化

目的：凸包体積空間及び予測誤差を最小にする。

50

1. 初期化：前回の出力データを、初期の凸包を規定する初期のクラスタ内に併合する。
2. i)凸包の外面の頂点；ii)凸包の適応可能な形状の移動境界の最小値及び最大値、を規定する。
3. 主ループ：モデルの誤差を特定する；ii)異常値を除外する；iii)シンプレックス幾何学的最適化を用いることによって凸包の新たな境界を規定する - 異常値を除去する毎に、境界を内側へ移動させる；iv)新たな凸包を計算する。異常値がそれ以上見つからずモデルの誤差が安定になるまで、このことを行う。
4. 出力：最適な凸包及び局所モデル予測。

【0059】

この手順の最後に、あらゆる新たなスペクトル x_i を予測することができるデータの最適な形状が得られることが期待される。数学的な詳細及びアルゴリズムの詳細はアルゴリズム 1 に提示する。

【0060】

以下は、特徴空間をマッピングして - 定量化値の知識ベースを構築することに関連する。同様な考え方に続いて、特徴空間全体にわたって上記自己学習プロセスを再帰的にマッピングすることができる。このマッピングは、ビッグデータであるスペクトルデータの特徴空間の大域的知識ベースを構成する。アルゴリズム 1 中の全部のステップを考慮に入れて、アルゴリズム 1 をこの特徴空間に再帰的に適用し、図 4 に示すステップ的な順序プロトコルがこれに続く。その手順は次の通りである：

目的：特徴空間 (4 a) 内の共分散の形状を順にマッピング (4 b) する。

1. 初期化：i)特徴空間の任意の所定点から開始する；ii)探索円の直径、探索方向の数、及び探索領域の次元を規定する；
2. アルゴリズム 1 を実行する：スペクトル x_i に対する凸包の局所的な直線形状を規定する。
3. 再帰的マッピング： x_i の最適化された凸包の内部で新たなデータ点 x_{i+1} を選択して、凸包を拡大するために抽出されるべき方向がそれ以上実現可能でなくなるまで、アルゴリズム 1 を再帰的に実行する。
4. 再サンプリング：特徴空間内で他に発見した位置に進む。
5. 主ループ：特徴空間の体積の所定割合をカバーすることが保証されるまで、動作 3 及び 4 を反復する。
6. 編集：キャッシュ記憶されるモデル (4 c) として使用されるモデルパスを登録することによって、特徴空間マップ内の知識ベースの定量化パスを編集する。
7. 出力：キャッシュ記憶されるモデル (4 c) の編集したマッピング。

【0061】

この手順の数学的詳細及びアルゴリズムの詳細は、アルゴリズム 2 に提示する。この再帰的マッピングの詳細は、図 17 b 中にも見出すことができる。

【0062】

このプロセスの結果は、図 4 c の特徴空間定量化値マップの構築である。このマップは、人工知能の方法及び装置の自己学習プロセスを構成する。マップ中の線はモデルの局所的予測のコヒーレントなパスを表し、即ち、新たなスペクトル X_{new} が、この線の凸包の付近に射影されると、この凸包内のその近傍点の変化のモードと同様なモードに従いやすく、この局所モデルのデータに基づいて予測することができる。この特徴空間の特性化は、次のことを可能にする：

- i. キャッシュ記憶されているモデルを用いて計算効率を加速すること - 新たなスペクトルが、前回の予測の線の凸包内に射影された場合、計算はキャッシュ記憶されているモデルで直接実行することができ、計算が直接的になる；
- ii. 定量化の異なるモードをもたらす代表的条件の特性化 - 多数の予測の線は、異なる種類の健康状態及びそれらの進展についての代謝情報を提供する；
- iii. これらの情報が特徴空間全体にわたってどれほど良好に表現されるかを判定すること - 十分なデータを有する領域のみが、正確な定量化及び効果的な探索を生成することを可

10

20

30

40

50

能にする；

iv. 時間軸に沿ったスペクトルパターンを理解するためのマップを提供すること、即ち、精密医薬の実現のためのスペクトルパターン認識の解釈を提供すること；

v. 教師なしのスペクトル情報を用いた状態診断用の高レベルの人工知能の基礎を提供して、複雑かつ多因子の非線形の分類マップの構築を可能にすること。

【0063】

以下は分類マッピングに関連する。分類のための自己学習人工知能の方法は、特徴空間内のクラスの形状を見出すという主要目的を有し、次のことによって見出す：i)当該クラスの局所的体積区間を最大にする；ii)非線形のクラスの場合に、特徴空間全体にわたる合計体積空間を最小にする；及びiii)関連する固有ベクトル変化によりクラス境界を区切ることで、クラス予測の誤差を最小にする。さらに、多数のクラスは高度に非線形にすることができ、特徴空間全体を通して極度に区分することができるものと期待することができる。多数のクラスは散乱したクラスを特徴空間全体にわたって有することもできる、というのは、他の状態が特徴空間変化において支配的であるからである。

【0064】

健康状態の複雑な分類に起因し、多数の状態が多変量であるので、教師付きクラスターリングは次のクラスに分類することができる：i)単一の一変量の診断 - 次元関数は単一パラメータの区間または閾値である；ii)排他的な一変量または多変量の診断 - 特徴空間内の各クラスのうちの極端な例のみが、他のクラスとのオーバーラップなしに識別される；及びiii)多変量の / 複合的な診断 - 特徴空間内の複数の状態からのデータのオーバーラップ部分のみを考慮に入れる（図5参照）。

【0065】

上記クラスターリングの基準は、複雑な健康状態を特性化してスペクトル空間にマッピングして、分類知識ベースを構成することを可能にする。分類知識マップを構築するための次の手順を開発した：

目的：クラスのロジスティック確率共分散の形状を特徴空間内に順次マッピングする。

1. 初期化：a) クラスターリング基準を規定する；a) 一変量クラス；b) 排他的クラス、またはc) 多変量クラス；及びd) 各クラスの閾値；b) 用意する：教師ベクトル s または行列 S 。

2. 凸包決定；i) 特徴空間内で教師データを選択する；ii) 特徴空間内の教師データの最大値及び最小値の座標を見出す；iii) 頂点のうちの1つを選択する；iv) 方向探索ボックスのサイズを規定する；iv) 凸包の体積空間増分基準 (v) を規定する。

3. クラスターの個性性をチェックする：上記最小値及び最大値がコヒーレントなクラス予測を保有することができない場合、クラスターを共通でない情報によって境界付け、最小の大域的な体積空間基準によって区分する必要がある、ここに $v_{optimal} = \max_{i=1}^n v_{local\ cluster} + \min v_{global}$ (最適な $v =$ (局所クラスターの v の $i=1$ から n までの総和の最大値) + (大域的 v の最小値)) であり；全クラスターが分離されている場合、 $\min v_{global} \max_{i=1}^n v_{local\ cluster}$ (大域的 v の最小値) (局所クラスターの v の $i=1$ から n までの総和の最大値)) であり；局所クラスターが1つしか存在しない場合： $v_{optimal} = \max_{i=1}^n v_{local\ cluster}$ (最適な $v =$ (局所クラスターの v の $i=1$ から n までの総和の最大値)) である；

4. 初期の探索：i) 固有ベクトルの数を決定して分類の誤差を予測する；ii) 統計的に不適正なデータを除外する；iii) 関連する方向が見出されない場合、 v だけ内側へ移動してステップ4を実行し；安定になるまでステップ1及び2を反復することによって、凸包の形状を再整形する。

5. クラスター境界を決定する：クラスター毎に、アルゴリズム2を実行し、ここでは教師ベクトル s 及び行列 S が、対応するスペクトルに関連するクラスター・クラスのロジスティック関数確率である。図5の、アルゴリズム2を用いてクラスター予測の凸包を見出す方法を参照。

6. 編集：特徴空間マップ内の知識ベースの分類クラスターを、これらのクラスターの凸包を登録することによって編集する。

10

20

30

40

50

7. 出力：キャッシュ記憶されるクラスタの編集されたマッピング。

【0066】

この手順の数学的詳細及びアルゴリズムの詳細は、アルゴリズム3中に提示する。

【0067】

この手順の最後に、特徴空間の完成したクラスタ・マップを分類知識ベースとして記録する。異なる状態についての全種類の分類の全合成物は、知識ベースの分類の複雑度を表し、状態間の相互作用、並びにこれらの状態の代謝的原因間の相互作用を学習することができる。新たなスペクトルを分類マップ内に射影することによって、対応する状態の予期される確率を、知識マップの座標に基づいて予測することができる。

【0068】

以下のことは局所的变化の形状に関連する。前節は、本発明のAIが、共分散固有ベクトル抽出を用いて、特徴空間全体にわたって探索アルゴリズムを反復することによって定量化及び分類を行って、定量化値及び分類のマップを提供することができる方法を説明した。变化の局所的形状の学習は、AIの発明の中心にある。

【0069】

スペクトルX及び組成データYの任意の集合は、それぞれ特徴空間F及びKに変換（例えば、カーネル、導関数、フーリエ、ウェーブレット、カーブレット）することができるものとする。FとKとの、及び

(外14)

T

(外14、以下「T」とする)

と

(外15)

U

(外15、以下「U」とする)

との、局所的潜在分散間の共分散を最大にするような、基底

(外16)

W

(外16、以下「W」とする)

及び

(外17)

C

(外17、以下「C」とする)

を見出さなければならない。問題は次の特徴空間内での局所的最適化に単純化され：

【数1】

$$f(w,c) = \operatorname{argmax}(t^t u)$$

($t^t u$ の最大値を与える引数)ここに、

(外18)

$$f = tw^t$$

10

20

30

40

50

及び

(外 19)

$$k = uc^t$$

であり：

(外 20)

$$w^t w = 1$$

10

及び

(外 21)

$$c^t c = 1.$$

に従う。ラグランジュ乗数法を適用して上記最適化問題を解くことによって、この式を再開して：

【数 2】

$$K^t F = W \Sigma C^t$$

20

とし、この式は

(外 22)

$$K^t F$$

(外 22、以下「 $K^t F$ 」とする)

の特異値分解であり、ここに

(外 23)

$$w = W[1,], \quad c = C[1,]$$

30

であり、関連する分散は

(外 24)

$$\Sigma[1,]$$

である。さらに、

(外 25)

$$F^t K K F^t w = \lambda w$$

40

及び

(外 26)

$$K^t F F^t K c = \lambda c.$$

と結論付けることができる。従って、

(外 27)

$$w$$

50

(外 27、以下「w」とする)

及び

(外 28)

c

(外 28、以下「c」とする)

は

(外 29)

$$\text{Cov}(\mathbf{F}, \mathbf{K})^2 = \text{Cov}(\mathbf{K}, \mathbf{F})^2$$

10

((F, K)の共分散の2乗 = (K, F)の共分散の2乗)の固有ベクトルであり、潜在空間

(外 30)

$\mathbf{t}^t \mathbf{u}$

(外 30、以下「 $\mathbf{t}^t \mathbf{u}$ 」とする)

$\mathbf{t}^t \mathbf{u}$ 内で表現され、ここにw及びcは共分散形状の特性次元を生じさせる。

【0070】

(外 31)

$$\mathbf{t} = \mathbf{u},$$

20

である際に：

(外 32)

$$f(\mathbf{w}, \mathbf{c}) = \text{argmax}(\mathbf{t}^t \mathbf{u})$$

(外 32、以下 $f(\mathbf{w}, \mathbf{c}) = \text{argmax}(\mathbf{t}^t \mathbf{u})$ とする)

($f(\mathbf{w}, \mathbf{c}) = (\mathbf{t}^t \mathbf{u})$ の最大値を与える引数)であるものと仮定すれば、同じ種類の導出が実現可能であり、特に有用である、というのは、収縮後に

(外 33)

t

30

(外 33、以下「t」とする)

が直交行列になるからである。この仮定は、他の固有ベクトル抽出アルゴリズムの基礎にもなる(非特許文献14)。

【0071】

$\mathbf{t}^t \mathbf{u}$ の形状を学習するためには、固有ベクトルw及びcの正規直交基底が必要であり、このため、局所的なF毎に、その局所的な特性次元及び形状を導出することができる。こうしたことはF及びKの収縮によって実現される：

【数3】

$$F_{i+1} = F_i - t_i w_i^t$$

$$K_{i+1} = K_i - u_i c_i^t$$

40

ここに：

50

(外 3 4)

$$t_i = F_i w_i, \quad u_i = K_i c_i,$$

及び

(外 3 5)

$$w_i = w_i / \|w_i\|, \quad c_i = c_i / \|c_i\|.$$

である。

10

F の最大の階数 (ランク) までの反復的収縮は、

(外 3 6)

t_i, w_i

及びそれらに対応する重要度を、固有ベクトル毎に捕捉した共分散

(外 3 7)

Σ

に関して解釈することによって、共分散の形状及びその複雑度を特定することを可能にする (非特許文献 1 5、非特許文献 1 6)。

20

【 0 0 7 2 】

t が直交である方法を用いる際に、収縮は次のように実行される：

ここに、

(外 3 8)

p

及び

(外 3 9)

q

30

は次式によって決定される：

【 数 4 】

$$F_{i+1} = F_i - t_i p_i^t$$

$$K_{i+1} = K_i - u_i q_i^t$$

$$p_i = F_i^t t_i (t_i^t t_i)^{-1}$$

$$q_i = K_i^t t_i (t_i^t t_i)^{-1}$$

40

【 0 0 7 3 】

(外 4 0)

P

(外 4 0、以下「 P 」とする)

と

(外 4 1)

50

Q,

(外 4 1、以下「Q」とする)
との関係より、直接的な線形モデルを
(外 4 2)

$$K = F\beta_{pls} + e$$

のように導出することができ、ここに：

【数 5】

$$\beta_{pls} = W(P^t W)^{-1} Q$$

ここに、
(外 4 3)

$$\beta_{pls}$$

(外 4 3、以下「 β_{pls} 」とする)
は PLS 回帰係数である。

【0074】

T の複雑な形状が斜めの射影 β_{pls} に凝縮され (非特許文献 2)、GLM が生成されることは、特に、非線形の特徴空間に起因して比較的大きな数の次元または成分を使用する場合に、ビッグデータにおける PLS の非効率の原因になる。従って、この戦略は、 $K^t F$ の局所的構造が、ほとんど、X に包含される Y についての体系的情報のみを有することを暗に意味し、分光信号の異なる尺度における変量効果をほとんど伴わない。さらに、適正な特徴空間変換は、

(外 4 4)

$$F^t F$$

(外 4 4、以下「 $F^t F$ 」とする)

及び $K^t F$ と同様な情報構造を得ることを可能にするものであり、このため理想的には：

【数 6】

$$(K^t K - \lambda_k) v_k = 0$$

$$(F^t F - \lambda_f) v_f = 0$$

$$(K^t F - \lambda_{kf}) v_{kf} = 0$$

局所的な最適化問題は $f(w, c) = \text{argmax}(t^t u)$ として提起されたままであるが、情報構造の理想的な制約は

(外 4 5)

$$(v_k \sim v_f \sim v_{kf}),$$

と同様である。完全な状態では、スペクトル情報は共線的な固有ベクトル構造を組成と共有して、干渉を無視できる純粋な化合物または物質により発生する。従って、第 1 成分における共変動の最大化を提供することが最重要である。

10

20

30

40

50

【 0 0 7 5 】

t 及び w の両者の解釈を A I に提供する方法は、両者を対にして直交させることである。直交する w を生成する P L S 定義を用いて t を直交させることができる：

【 数 7 】

$$F = TW^t; T = USV^t$$

$$F = USV^t W^t = US(V^t W^t)$$

$$F = T_o W_o^t$$

10

ここに、
(外 4 6)

$$T_o = US$$

及び

(外 4 7)

$$W_o^t = (WV)^t$$

20

である (非特許文献 1 7、非特許文献 1 8)。

【 数 8 】

$$T_w = T(P^t W)^{-1}$$

によって、直交スコア (得点)

(外 4 8)

$$T_w$$

30

(外 4 8、以下「 T_w 」とする)

への直接的対応が存在する。

従って、A I は、対応する垂直な W と直交する T_w または

(外 4 9)

$$T_o$$

(外 4 9、以下「 T_o 」とする)

におけるパターン分析を実行して、局所的な物質及びモデルのコヒーレンスを導出する方法を有する。

40

【 0 0 7 6 】

次の特徴空間の内部的関係：

【 数 9 】

$$t_i = T_w \beta_T$$

を尊重し、この関係は次式の最小二乗推定値によって与えられる：

【 数 1 0 】

$$\beta_T = (T_w^t T_w)^{-1} T_w^t t_i$$

50

【 0 0 7 7 】

これにより、あらゆる特徴空間サンプルの T_w への射影が、関連する局所的な固有構造に対するコヒーレントな線形干渉パターンに従う。従って、 T_w に射影されるあらゆる所定の新たなデータが

(外 5 0)

$$t_i = T_w \beta_T.$$

なる信頼区間内に包含される。

【 0 0 7 8 】

局所的なデータ集合（データセット）の複雑度は、図 6 に示すようにサンプル及び変数を共に改良することによって低減することができる。アルゴリズム 1 の局所的な方向選択においては、整合性のある固有構造を提供するサンプル及び変数のグループを見出さなければならない。従って、残差の交差検証（相互検証）の使用とは逆に、共分散の局所的安定性を表す適合度関数も、最適化手順として提案する。局所的最適化問題は次の性質を有する：i) 開始時のデータ集合の数を盲目的に選択する；ii) 関連する PLS 回帰をデータ集合毎に実行する；iii) スコア空間内に射影する

(外 5 1)

($T=FW$)

iv) ロバスト（頑健）な線形回帰を用いて T 内部の固有構造を識別する（例えば、RANSAAC: random sample consensus ランダム・サンプリング（標本抽出）による合意）；v) ある閾値に達するまで、この手順を再実行する。しかし、 T スコア空間内の線形モデルの存在は、収縮が、局所的な線形 PLS モデルを構築するために用いる全データ内での体系的変化をモデル化していることを意味する。

【 0 0 7 9 】

各共分散の方向の整合性は、交差検証（例えば、 n 個抜き）によって決定され、ここでは全データ点が局所的な固有構造を維持しなければならない。あらゆる新たな未知のデータについて、予測を次のように実行する：

i) 予測の整合性を、選択した部分空間内で決定する；ii) 全部の学習セットの低いバイアス分散；iii) 選択したデータ集合内の異なるデータを用いた、未知のデータの予測の低いバイアス分散；iv) T の固有構造の線形性の存在。さらに、あらゆる未知のデータの予測可能性は、 T 内で抽出した線形固有構造の信頼区間を導出することによって得ることができ、これにより、予測の p 値も予想することができる。所定の閾値を上回る p 値は、明確であるか予測不能であるものと考えることができる。上記 AI は、予測が必要な正確性を有するか否かを「先験的に」知ることができる、というのは、周知のコヒーレントな固有構造のみを用いて予測を実行するからである。

【 0 0 8 0 】

適正な局所的最適化モデルを提供するためには、次のことをしなければならない：

i) 残差及びその収縮構造を最小にする；ii) サンプルの数を最小にして、同じ共分散内の変数の個数を最小にする；iii) T 空間内部のコヒーレンスを保証する；及び iv) F と K との間の固有構造の類似性。これらの目標の全部を考慮に入れて、これらを次式的最適化関数で表現することができる：

【数 1 1】

$$J = \operatorname{argmin} \left(\operatorname{PRESS} \times \left[\frac{n_{pc}}{n} + \frac{n_{sel.vars}}{n_{vars}} + \frac{1}{\operatorname{cov}(V, W)} + p\text{-値} \right] \right)$$

10

20

30

40

50

この式は「補正した」PRESS (Predicted Error Sum of Squares: 予測誤差の平方和、予測平方和) と考えることができ、ここに n_{pc} は成分の数であり、 n はデータの数であり、 $n_{sel,vars}$ は変数の全体 n_{vars} から選択した F の変数の個数であり、
(外52)

$$n_{vars}, cov(V,W)$$

は $F^t F$ の固有ベクトルと $K^t F$ の固有ベクトルとの共分散であり、 p 値は T 空間内の最小二乗モデルの確率値である。この最適化の最後に、最適な局所モデルが得られる。別なモデルの改良は、直交情報フィルタリング (フィルタ処理) によって実行される。

10

【0081】

以下は、局所的部分空間のコヒーレンスに関連する。局所の特徴空間 F のコヒーレンスは：
i) F と K との固有構造の類似性；
ii) 低い複雑度；
及び iii) 情報決定によって保証される。
 F と K とは：

【数12】

$$j = \operatorname{argmax} (V_k^t V_F)$$

$$\text{ここに: } F = U_F S_F V_F^t = T_F V_F^t; \text{ かつ } K = U_K S_K V_K^t = T_K V_K^t.$$

20

である際に、同様な固有構造を有する。

【0082】

X 及び Y の効率的な変換が存在し、ここに
(外53)

$$F = f(X)$$

及び

(外54)

$$K = f(Y)$$

30

であり、このため理想的には

(外55)

$$T_K = T_F.$$

である。スペクトル情報はマルチスケールであるので、次の信号ベース (例えば、フーリエ、ウェーブレット、カーブレット) のマルチスケール最適化を提案することができる：

40

【数13】

$$F = \sum_{i=1}^2 \theta_i \mu_i$$

【0083】

ここに、 i は選択した個別の信号尺度であり、これにより
(外56)

$$V_k^t V_F$$

50

が最大にされる。

【 0 0 8 4 】

$K^T F$ の固有構造は極めて重要なものであることが好ましい。F の複雑度は、その固有値の分布によって推定することができ、固有値 は特徴空間の特徴的な次元を規定する。分光法の信号では、 が次の限界値まで指数関数的に減少することが期待される：

【数 1 4】

$$\Sigma_i = \Sigma_r + (\Sigma_1 - \Sigma_r) e^{-ki}$$

ここに、
(外 5 7)

Σ_i

10

は期待される i 番目の固有値であり、
(外 5 8)

Σ_r

は残差の固有値であり、
(外 5 9)

Σ_1

20

は最大の固有値であり、
(外 6 0)

k

は減衰係数である。特徴空間の局所的複雑度は、次の測定基準によって測定することができる：

【数 1 5】

$$C = npc / (k n)$$

【 0 0 8 5 】

(外 6 1)

$k \rightarrow +\infty$

40

である際に、
(外 6 2)

$C \rightarrow 0$.

を暗に意味する。
(外 6 3)

$npc \rightarrow 1$

50

かつ
 (外 6 4)
 $n \gg npc.$

である際に、こうした限界は漸近的に近似される。

(外 6 5)
 $\Sigma_r \rightarrow 0,$

10

である際に、 $K^t F$ は階数不足 (階数落ち) である。固有構造のランダム性は、 F 、 K 及び $K^t F$ のランダム化によって得られる (非特許文献 19)。行のランダム化は、行ベクトルを生じさせる固有ベクトルの数を定める試料スペクトルの限界を決定することを可能にするのに対し、列のランダム化は、変数が列ベクトルを生じさせることを可能にする固有ベクトルの限界を定める。固有値の数の統計的安定性は交差検証によって提供される。

【 0 0 8 6 】

以下は部分空間情報の最適化に関連する。前の手順が与えられると、特徴空間の選択した方向は安定な線形モデルを既に提供している。 Y を予測するために最小数の固有ベクトルしか必要としないことがさらに期待される。信号が前処理 (例えば、ベースライン、散乱効果、迷光) されて、より良好な特徴空間の基底に変換されているにもかかわらず、体系的な干渉がデータ中に常に存在する。こうした干渉は、第 1 成分を超えるスコア・ローディング

20

(外 6 6)
 (t-p)

関係に悪影響を与える。従って、理想的には、関係は 1 つの固有ベクトルだけで得るべきである。こうしたことは必ずしも実現可能ではないが、直交フィルタリングによってモデルの関係を大幅に簡略化することができる。

【 0 0 8 7 】

30

F と K とは、互いに関係しない体系的情報を有することができる。従って、局所的情報を構造化する方法、即ち F と K とが共通に保有する情報の量、及び独立した情報の量を、直交フィルタリングを実行することによって知らなければならない (Trigg 及び Wold : 2002、Trygg 及び Wold : 2003、Bylesjoetal : 2008)。図 7 に、直交フィルタリングがより低い複雑度の固有構造を生じさせる様子を示す。図 7 a ~ 7 c は、前のステップが、

(外 6 7)

X_1, X_2

(外 6 7、以下「 X_1 、 X_2 」とする)

40

及び
 (外 6 8)

X_3

(外 6 8、以下「 X_3 」とする)

なる部分集合を用いて局所的な較正データ集合を最適化した様子を示し、上記線形モデルの複雑度を 3 つの予測潜在変数のみまでに大幅に低減している。こうした結果は、 X_1 、 X_2 及び X_3 なる部分集合が、体系的な干渉を有し、直交フィルタリングを施すことができ、これにより：

50

【数 1 6】

$$F=TP^t+T_0P_0^t$$

$$K=TQ^t+U_0P_0^t$$

となり、ここに T は、F と K との間に共通する情報のスコアであり、共通する情報は共分散を最大にし、P、Q は対応するローディングである。T₀ 及び

(外 6 9)

U₀

10

(外 6 9、以下「U₀」とする)

は共分散に対する直交スコアであり、

(外 7 0)

P₀, Q₀

は対応する直交ローディングであり、即ち、T₀ は K に直交し、U₀ は F に直交する (Trigg 及び Wold : 2002)

20

【0 0 8 8】

サンプル及び変数の再帰的選択によって、

(外 7 1)

TP^t

を最大にすることができ (図 7 c ~ 7 d)、潜在変数の個数は、1 に近い最適より低いレベルまで最小にされ、即ち、収縮を必要とせず、F と K との間に直接的な対応が存在する。適正な特徴空間変換が

(外 7 2)

30

$$T_0P_0^t \rightarrow 0$$

をもたらし、

(外 7 3)

$$F=TP^t$$

は通常の PLS アルゴリズムによって得られることが期待される

【0 0 8 9】

40

(外 7 4)

U₀P₀^t

同様に、

(外 7 4、以下「U₀P₀^t」とする)

は 0 になる。分析グレード品質を伴うあらゆる定量化値は、その定量化値に直交するあらゆる体系的变化を有するべきでない。U₀P₀^t が大幅である際には、AI を適切に学習させて正確な予測を提供することができないことを意味する、というのは、元の学習情報に体系的誤差の問題があるからである。適正な条件下では、

50

(外 7 5)

$$U_o P_o^t \rightarrow 0$$

及び

(外 7 6)

$$T P^t \gg T_o P_o^t$$

となり、このため

(外 7 7)

$$K = F \beta_{pls.}$$

となる。

【 0 0 9 0 】

$T_o P_o^t$ が大幅である場合、特徴空間変換が非効率的になる。これらの場合、予測は直交フィルタをまず適用することによって実行され、

(外 7 8)

$$t_o = f p_o / (p_o^t p_o),$$

及び

(外 7 9)

$$f_{corr} = f - t_o p_o^t$$

及び

(外 8 0)

$$k = f_{corr} \beta_{pls.}$$

となる。完全にするために、上記方法はアルゴリズム 5 中に記述する (図 1 7 c 参照) 。

【 0 0 9 1 】

以下は部分空間特性化用の測定基準に関連する。提案する方法の主な利点の 1 つは、局所的学習測定基準のマップを含めることによる自己学習知識ベースの特性化を可能にすることであり、局所的学習測定基準は、例えば i) データ表現の数 ; ii) 固有構造の複雑度 ; iii) F と K との共直線性 ; iv) 予測される平方の総和 ($P R E S S$) ; v) $K^t F$ の分散 ; 及び vi) モデル情報構造である。知識ベースの詳細な測定基準は表 1 中に提示する。

【 0 0 9 2 】

特徴空間を特性化することによって、AI システムは、特徴空間の異なる領域が定量化及び定性化をカバーする様子を正確に知ることによって、自己学習及び予測を共に管理する。

【 0 0 9 3 】

以下は自己学習機構に関連する。前節は、自己学習方法及び装置のアルゴリズム及び数学的手順を例示している。ここでは、上記の手順を、人間の介入なしに供給データから自己学習を自動実現するシステムにまとめ上げる方法を提供し、これにより、このシステムは次のことができる : i) ビッグデータの分光データにおける 0 のデータから大量のデータまでのデータ供給によって自律的に学習すること ; ii) 共分散を最良に捕捉する最良のマルチスケール特徴空間を決定すること ; iii) 知識ベースに基づいて新たな未知のデータを予測

10

20

30

40

50

し、予測不可能なデータを取り扱う方法を予測すること；iv)定量化値及び分類のマップを自己学習で構築し、これらのマップを用いて、新たなデータからのコンピュータによる計算効率の良い予測及び学習を実行すること。

【0094】

体液及び体内組織における生物学的変動性は極めて膨大である。ビッグデータでは、あり得るすべてのスペクトルの組合せに対処するモノリシックなモデル戦略を構築することができるロバストな知識ベースを構築するための代表的な試料の意味を決定することは決してないであろう。さらに、生体系は進化し、生体系の生化学は常に変化しつつあり、新たな細胞、新たなタンパク質、新たな代謝産物を生じさせる。従って、生体系に適用される分光法のAIは常に自己学習しなければならない。開発したシステムは、このシステムが予測することができない新たなデータを絶えず追加することによって、初期の、非常に限られた知識ベースから自己学習することができる。このシステムは、前節の測定基準及び方法を用いることにより、特徴空間及び初期の知識ベースを計算することによって動作を開始する。特徴空間の各部分空間の予測可能性を管理することによって、新たに取得したデータが予測可能であるか、学習サイクルに入るべきであるかのいずれかである場合に、システムは持続することができる。予測することができない場合には、近傍点を有さない（例えば、動作の開始時には、どのシステムも決して特徴空間全体をカバーしない）か、整合性のあるモデル化を有さないかのいずれかであるデータの機密保存場所として機能する検査データベースにデータを追加する。一旦、新たに収集したデータが完結すると、検査データベース内に収集したデータは知識ベースのみに渡されて、特徴空間の対応する領域が、コヒーレントな部分空間知識ベースの開発を可能にする。

10

20

【0095】

図8に、自己学習プロセスの主な機構を示す。上記システムは、初期には限られた数のスペクトル対を供給され、これらのスペクトル対は組成X及びYに対応するものとする。何らかの新たなXが記録されると、この新たなXを初期の特徴空間マップ内に射影して、既存の知識ベースに属するか否かをテストする。この射影が既存のモデルパスの付近にある場合、キャッシュ記憶されている既存のモデルを用いて予測を策定する。この予測が期待される品質の範囲内でない場合、かつ期待される品質の範囲内であることを可能にする近傍点が存在する場合、新たなモデルをアルゴリズム1によって構築し、予測を実行して、アルゴリズム2によって得られた対応するモデル及びパスをキャッシュ記憶する。

30

【0096】

何らかの新たなスペクトルが特徴空間内に射影され、近傍点を有さない際に、この新たなスペクトルを直ちに検査する。システムは学習サイクルに入り、検査すべき試料の組成を提供するように、ユーザまたはシステムに求める。一旦、組成がXとYの対を有すると、システムは検査の近傍点を探索する。組成が近傍点を有さない場合、データを単に検査中のままにする。組成が近傍点を有する場合、アルゴリズム1及び2を用いて学習プロセスを開始して、両方の局所モデルを探索して局所的な共分散マップを構築する。新たなデータが、検査中のデータと共に、整合性のある局所モデル及びモデルパスを生成することができる際のみ、このデータを知識ベースに渡すことを認定する。知識ベースは、新たなデータが追加されると、絶えず更新を受けて、予測は特徴空間の新たな領域まで拡張される。

40

【0097】

この意味で、上記システムは：i)知識ベース内にない予測は決して生成せず；ii)検査データベースを維持して学習し；iii)認定されたデータベース内へ検査データを渡すことを認め；iv)認定されたデータのみを用いて知識ベース及び予測を構築し；v)人間の介入なしに自己学習し；iv)データのサイズとは無関係に、供給されたデータで知識ベースを成長させる。さらに、この方法は、知識ベースを構築して予測を実行することを開始するために、深層学習ニューラル・ネットワークのように大規模なデータベースを必要としない。上記システムは認定された知識ベースのみを使用し、従って、他のモデル化方法のように、予測にバイアスの問題がない、というのは、他のモデル化方法は、大域的に安定したモデル

50

・アーキテクチャを生成するために大量のデータを必要とするからである。共分散マップ、分類マップ、及びキャッシュ記憶されているモデルは、システムを、コンピュータによる計算効率が非常に良いものにする。上記システムは、数学モデルを構築するために、今日の従来技術のシステムのような人間の介入を必要としない。

【 0 0 9 8 】

XをFに変換しYをKに変換する変換の適正な基底を見出すことは、局所的線形モデルが第V節に提示するように抽出される広範囲の特徴空間を構築することの中心にある。特徴空間変換のための基本原理は、FとKとの間の固有構造の類似性の最大化である。基底変換が、XとYとの間とは無関係な体系的分散及びノイズをフィルタ処理することができる場合、FとKとは固有構造が等しくなる。

10

【 0 0 9 9 】

図9に、特徴空間変換を実行する方法を示す。あらゆる分光信号が直交基底（例えば、フーリエ、ウェーブレット、カーブレット）に分解される。これらの基底は、当該規定の性質に基づいて信号の尺度を再構成するための独立した基底を提供する。存在すれば、あらゆる代謝産物についての情報を、スペクトルの異なる尺度にわたって分散させ、従って、特定分子にとって最適なスペクトル変化を、尺度再構成を用いて元の信号から抽出しなければならない。

【 0 1 0 0 】

フルスペクトルの分解後に、固有構造の類似性をFとKとの間で最大にする組合せを提供する最適な規定を見出さなければならず、ここに

20

(外81)

$$F=T_f V_f^t$$

かつ

(外82)

$$K=T_k V_k^t$$

であり、

(外83)

$$V_k^t V_f$$

30

は最大になる。情報の完全な整合下で、

(外84)

$$T_f=T_k=T$$

であり、このためF及びKは共に同じ固有構造を有する。なお、N I P A L S (nonlinear iterative partial least squares: 非線形反復部分最小二乗) P L S、相関を最大にする前の仮定の下では、情報の当該部分中の変換されていないXとYとが同じ固有構造を有し、即ち、同じスコアTを有する。ここでは、まずスコアの類似性の仮定を可能にする特徴空間をまず構築し、このことは提示する発明の成功に大きく寄与する。一旦、特徴空間変換が実現されると、直接の線形(直線的)関係がKとFとの間に存在することが期待され、

40

(外85)

$$K=F\beta.$$

50

となる。従って、発明者が証明するものは、PLSまたはSVM型の仮定は、KとFの固有構造が同様である場合のみに可能である、ということである。さもなければ、体系的情報が、スコア、内的関係の仮定、及びコヒーレンスの質を低下させる。

【0101】

従って、K及びFを共に直交基底

(外86)

μ

(外86、以下「 μ 」と称する)

10

に分解することを考える：

【数17】

$$F = U_f \mu$$

$$K = U_k \mu$$

ここに、

(外87)

$$U_f = U_k^t$$

20

の誤差

(外88)

e

を最小にする

(外89)

U_f

30

(外89、以下「 U_f 」とする)

と

(外90)

U_k

(外90、以下「 U_k 」とする)

との組合せが存在し、

(外91)

$$\beta = (U_f^t U_f)^{-1} U_f^t U_k$$

40

である。固有構造の類似性の最大値を見出す問題は、FとKとの間に共通する情報を最大にする、 U_f と U_k との最良の線形結合を見出す最適化問題であり、特徴空間変換を自律的に規定する。

【0102】

この変換を実行することによって、スペクトル及び組成の無関係で体系的かつランダムな成分の大部分が除外される。上記システムは、シンプレックス・アルゴリズム、粒子群最適化アルゴリズム、及び遺伝アルゴリズムのような進化アルゴリズムによって特定の代謝産物を定量化する μ の最良の組合せを抽出する方法を自己学習する。一旦、特徴空間変

50

換を特定の部分空間について学習すると、上記システムは、再計算する必要がなくなり、この変換を直接用いて予測を生成する。

【 0 1 0 3 】

図 1 0 は、特徴空間変換のフルクソグラムを示し、ここでは：i)元の信号を分解し；ii)最良の基底の初期の推定値を線形回帰によって推定し；そしてiii)これらの基底の組合せを進化的方法によって最適化する。固有構造の基準を満足するような規定の組合せを見出した場合、変換についての情報をキャッシュ記憶して、将来の予測において特徴空間を構築するための特徴空間変換として使用する。

【 0 1 0 4 】

以下はキャッシュ記憶されているモデル、共分散マップ及び分類マップに関連する。キャッシュ記憶されているモデル、共分散マップ及び分類マップを用いることは、コンピュータの計算効率、自己学習する人工知能にとって最重要であり、コンピュータ・リソース（資源）の大幅な節約をもたらす。図 1 1 a は、キャッシュ記憶されているモデルを用いて予測を加速する方法を示す。一旦、新たなスペクトルが記録されると、この新たなスペクトルを特徴空間内に射影して、その近隣のモデルパスをチェックする。近隣のモデルパスがあれば、予測は第 I V 節の方法を用いて実行し；一旦、何らかの新たなスペクトルが記録されると、次の動作を実行する：

i)キャッシュ記憶されているモデルが予測を正確に実行することができる場合、その結果をエンドユーザに対して提示する；

ii)近隣のモデルが品質予測の境界閾値を提示することができる場合、上記システムは、新たなモデルを計算して知識ベースを更新する前に合意の予測を提供することができる；及び

iii)近隣のモデルが十分な品質予測を提供しない場合、局所モデルの新たな探索を実行して、新たなモデルパスを知識ベース内に展開する。

【 0 1 0 5 】

以下は結果及び考察、特に定量化に関連する。ここでは、上記の自己学習人工知能法の有効性の実証を、未知の血液及び血清の予測をベンチマーク（比較基準）とすることによって実行する。結果を、最新技術のケモメトリクスの部分的最小二乗回帰（PLS）大域モデルと比較して、従来技術との比較の単純な基礎を提供する。この大域的 PLS は、交差検証によりバイアス分散間のバランスをとって、最小数の固有ベクトルまたは潜在変数（LV：latent variables）を導出することによって得た。血液及び血清の未知の試料の全体を：i)モデルの複雑度；ii)予測の平均誤差(%)；及びiii)共直線性 - ピアソン（Pearson）相関（ R^2 ）に関して分析した。

【 0 1 0 6 】

図 1 3 に、PLS が血液のような生体液の複雑性に対処することができない理由を例示する。赤血球が血液の主要な細胞成分であり、ヘモグロビンの含有量に直接関係するにもかかわらず、赤血球細胞の量を正確に予測するには、線形モデルで十分であることを期待することができた。図 1 3 a は、まさにその逆を示し、赤血球スペクトルの定量化が、大きな干渉によって影響される高度な非線形であることを示し、このため、PLS モデルは、高い赤血球カウント数（例えば、 $> 5 \times 10^{12}$ 細胞/L）において、非常に高い分散及び大幅なバイアスを示している。干渉は PLS モデルの 7 つの LV によって表現される。この大きな分散は、主要な成分でも複雑なスペクトルパターンを呈すること、一旦、線形定量化に単純化すると大幅な予測バイアス（11.50%、表 1）が得られることを意味する。一般的な線形モデルは、ヘルスケアにおける分析グレード予測を実現するために苦労する。

【 0 1 0 7 】

図 1 3 c は、白血球についての PLS 予測を示す。白血球は血中に赤血球よりも低い濃度で存在するが、それでも細胞成分のうちの大きな割合である。大きさの差は、白血球を PLS で予測することができないことを示すのに十分である。図 1 3 c の結果は、予測が非常に大幅な分散及び大きなバイアスを有することを示している。PLS は、大きな個数の LV（10 個）で 27% の誤差（ $R^2 = 0.45$ ）を有するモデルを提供することしかで

きず、相当量のスペクトル干渉が血液全体中の白血球の定量化に悪影響を与えることを示している。

【0108】

赤血球及び白血球は、上記の自己学習方法がスペクトル情報の複雑性に対処して、局所的なマルチスケールモデル化に基づいて正確な予測を提供する方法についての好例である。図13b及び13dは、それぞれ赤血球及び白血球についての自己学習人工知能の結果を提示する。

【0109】

両パラメータが非常に低い分散及びバイアスを示して、診断のための医療グレードの定量化を可能にし、2.4%及び5.15%の誤差しかなく、非常に大きな相関を有する(表1)。

10

【0110】

最も重要なことは、両モデルにおいて1つのLVまで複雑度を低減していることである。この自己学習人工知能は、局所的なマルチスケールの線形関係、フィルタ変数、及びサンプルを見出すことができ、これにより、スペクトル情報と定量化値との直接の関係が見出され、生体試料における複雑な干渉の影響をフィルタ処理で除外している。

【0111】

表1は、血液及び血清全体のパラメータについての定量化結果を要約する。これらのパラメータは、赤血球、ヘモグロビン、ヘマトクリット値、MCV (mean corpuscular volume: 平均赤血球容積)、MCHC (mean corpuscular hemoglobin concentration: 平均赤血球ヘモグロビン濃度)、白血球、及び血小板のようなヘモグラム(血液像)パラメータである。結果は、上記の自己学習方法及び装置による非常に大幅な改善を示し、ここでは全パラメータの推定値が、学習した範囲内で6%を下回る誤差を呈している。

20

【0112】

図14に、血清中のビリルビン及びミオグロビンについての結果を示す。ビリルビンは血清中の重要な成分であり、黄色-茶色の彩色を有する。ミオグロビンはより低い量で存在するが、存在する際には、そのスペクトル指紋は、可視-近赤外領域では血清中で非常に大きい。従って、両分子をPLSモデルによって直線的に定量化することができることも期待される。図14a及び14c中の結果は、ビリルビン及びミオグロビンのPLS予測が非常に大幅な分散を呈し、それぞれ12.5%及び31.0%の誤差を伴うことを示している。これらの分子がスペクトル信号中で非常に強い指紋を与えるにもかかわらず、これらの分子にはまだ大きな干渉の問題がある。

30

【0113】

最も関係のある結果は、図14b及び14dに開示する自己学習人工知能の方法を用いる際に、バイアス-分散が大幅に低減されることである。大部分のモデルは複雑度を減少させ、より高い濃度で提示されるすべてのパラメータが、1つの固有ベクトルの射影(1つのLV)しか使用しない。提案する方法は、分子の定量化値と線形相関のある局所的なマルチスケールのスペクトル情報を見出すことができた。この意味で、学習した全パラメータ(赤血球、ヘモグロビン、ヘマトクリット、MCV、MCHC、白血球、及び血小板)が、分析グレード品質を6%未満のバイアスで達成することができた。

40

【0114】

血清についても同様な結論が得られ、ここではビリルビンのような高濃度のパラメータ、あるいはミオグロビンのような高い吸光度のパラメータが、1つのLVのみを用いて直接、定量化可能である。グルコース(ブドウ糖)、クレアチニン、CRP (C-reactive protein: C反応性タンパク)、トリグリデド、尿素、及び尿酸のような他のより低濃度のパラメータは、当該パラメータのモデルの複雑度を2~3個のLVまで大幅に低減した。こうしたことは、より低濃度のパラメータには、より大きな干渉及び局所的变化の問題があること、並びにこれらのパラメータの精度が検出器の背景ノイズによって影響され始めることの現れである。

【0115】

50

図15に、PLSのベンチマーク対本発明の自己学習人工知能を提示する。PLSモデル化は：赤血球、ヘモグロビン、MCV、MCHC、血小板、ビリルビン、及びCRPについてのPOC（point of care：臨床現場即時）の定性的な定量化を支えるに過ぎない。これらのパラメータの誤差はおよそ7%～12%である。PLSモデル化を用いて推定した他のすべてのパラメータは、POCについての15%の誤差基準を満足しなかった（図15参照）。自己学習AIは、次のパラメータ：赤血球、ヘモグロビン、ヘマトクリット、MCV、MCHC、白血球、血小板、ビリルビン、グルコース、ミオグロビン、CRP、トリグリセリド、及び尿酸において、医療の分析グレード品質を実現することができた。クレアチニン及び尿素の定量化値のみが5%の限界を上回ったが、POC定性分析用には確かに適格とされた。提案する自己学習人工知能の方法は、背景技術において提示した以前の技術上の障害を大幅に解決し、分光法が分析級の誤差を実現することを可能にした。

10

【0116】

以下は結果及び考察、時に分類に関連する。ここでは、提案する自己学習方法の有効性を：貧血症、白血球増加症、血栓性血小板減少紫斑病、血小板増加症、肝不全、真正糖尿病、急性心筋梗塞、腎機能障害、及び炎症のような既知の健康状態の分類についても実証する。これらの状態の分類は、次の診断カットオフ値に応じて実行した：i)貧血症 - $4 \times 10^{12}/L$ を下回る赤血球カウントレベル及び 13 g/dL を下回るヘモグロビンレベル；ii)白血球増加症 - $10^{10}/L$ を上回る白血球レベル；iii)血栓性血小板減少紫斑病 - $100 \times 10^9/L$ を下回る血小板レベル；iv)血小板増加症 - $400 \times 10^9/L$ を上回る血小板レベル；v)肝不全 - 1.2 mg/dl を上回るビリルビンレベル；vi)真正糖尿病 - 100 mg/dL を上回るグルコースレベル；vii)急性心筋梗塞 - 147 ng/ml を上回るミオグロビンレベル；viii)腎機能障害 - 1.3 mg/ml を上回るクレアチニンレベル；ix)炎症 - 2.0 mg/dl を上回るC - 反応性タンパクレベル。

20

【0117】

表2は、提示された状態についての分類結果を、真及び偽に関して、それぞれ陽性及び陰性との組合せで提示する。この結果は、自己学習の分類が線形分類装置、ロジスティックPLSよりも優れていることを示している。このことは、診断用のカットオフ値が、血栓性血小板減少紫斑病用、あるいは高レベルの白血球を有する感染症（白血球増加症）のような複雑な干渉の問題がある状態用のような低濃度におけるものである状態にとって特に重要である。大域的PLSモデルは、貧血症、血小板増加症、及び急性心筋梗塞のための臨床現場即時医療を（15%の分類誤りで）支えることができるに過ぎない。大部分のパラメータは、正確な診断の機会を50%～80%のレベルで呈し、従って、線形分類を用いることは、健康状態の分類にとって非常に限られることが判明する。

30

【0118】

自己学習方法は、85%を上回る正確な診断の機会を常に奏することができた。自己学習方法は、貧血症、血小板増加症、及び急性心不全の場合の100%を正確に診断することができた。白血球増加症、真正糖尿病、及び肝機能のような状態も、完全に近い分類（97%の正確である機会）を実現している。こうしたことは、誤分類される値がカットオフに近く、この分類方法では検査過誤を考慮に入れなかったからである。検査過誤を考慮に入れれば、5%の許容誤差（エラーマージン）で、これらの状態も100%分類される。血栓性血小板減少紫斑病及び腎機能障害は、それぞれ87%及び89%の分類率を有する（表2参照）。こうした結果は期待されていた、というのは、血小板及びクレアチニンの値は、スペクトル中のそれらの信号情報のためには大幅に低いからである（例えば、クレアチニンは自己学習を用いれば14%の予測を有する、表1参照）。それにもかかわらず、上記2つの状態は15%を下回る分類誤りである。

40

【0119】

本明細書中に特に断りのない限り、記載した文章または流れ図中のステップの特定の順序は例示に過ぎず、本発明から逸脱することなしに変更することができることは、通常の当業者にとって明らかである。従って、特に断りのない限り、記載したステップは順序不

50

定であり、可能であれば、これらのステップは任意の好都合な、あるいは望ましい順序で実行することができる。

【 0 1 2 0 】

本明細書中に説明した本発明の特定の実施形態は、ファームウェア内及び/またはコンピュータで使用可能な媒体上に存在するコード（例えば、ソフトウェア・アルゴリズムまたはプログラム）として内蔵させることができることは明らかであり、このコードは、本明細書中に記載したいずれかのサーバーのような、コンピュータ・プロセッサを有するコンピュータシステム上での実行を可能にする制御論理を有する。こうしたコンピュータシステムは一般に、上記コードの実行による出力を提供するように構成されたメモリ記憶装置を含み、このコードは、この実行に従ってプロセッサを設定する。このコードは、ファームウェアまたはソフトウェアとして構成することができ、一組のモジュールとして編成することができ、これらのモジュールは、ディスクリットな（個別の）コードモジュール、関数呼び出し、手順呼び出し、あるいはオブジェクト指向プログラミング環境ではオブジェクトのような、本明細書中に説明した種々のモジュール及びアルゴリズムを含む。モジュールを用いて実現される場合、このコードは単一のモジュールまたは複数のモジュールを含むことができ、これらのモジュールは互いに協働してマシンを構成し、このマシンにおいて上記コードが実行されて、本明細書中に説明する関連する機能を実行する。

10

【 0 1 2 1 】

本発明は、形はどうであれ説明した実施形態に限定されるものと見るべきでなく、通常の当業者は、その変形の多数の可能性を予見する。上述した実施形態は組み合わせ可能である。以下の特許請求の範囲は、本発明の特定の実施形態をさらに提示する。

20

【 0 1 2 2 】

参考文献

P. Geladi and B. Kowalsky, "Partial least squares regression: a tutorial", *Analytical Chemical Acta*, 185:1-17, 1986 (非特許文献 1)

A. Phatak and S. Jong, "The geometry of partial least squares", *Journal of Chemometrics*, 11:311-338, 1997 (非特許文献 2)

Huang, G.B. Huan, S. Song, and K. You, "Trends in extreme learning machines: A review", *Neural Networks*, 61:32-48, 2013 (非特許文献 3)

L. Ramirez-Lopez, T. Behrens, K. Schmidt, A. Stevens, J.A.M. Dematte, and T. Scholten, "The spectrum-based learner: A new local approach for modeling soil vis-nir spectra of complex datasets", *Geoderma*, 195-196:268-279, 2013 (非特許文献 4)

30

D.P. Solomatine, Maskey. M., and Shrestha. D.L., "Instance-based learning compared to other data-driven methods in hydrological forecasting", *Hydrological Processes*, 22:275-287, 2008 (非特許文献 5)

T. Neas, T. Isaksson, and B. Kowalski, "Locally weighted regression and scatter correction for near-infrared reflectance data", *Anal. Chem.*, 62(7):664-673, 1990 (非特許文献 6)

C.D. Christy and S.A. Dyer, "Estimation of soil properties using a combination of spectral and scalar sensor data" (非特許文献 7)

40

J.S. Shenk, M.O. Westerhaus, and P. Erzaghi, "Local prediction with near infrared multi-product databases", *Journal of Near Infrared Spectroscopy*, 5:223-232, 1997 (非特許文献 8)

T. Fearn and A.M.C. Davies, "Locally-biased regression", *Journal of Near Infrared*, 11(6):467-478, 2003 (非特許文献 9)

A.M.C. Davies and T. Earn, "Quantitative analysis via near infrared databases: compression analysis using restructured nearest infrared and constituent data-deux (carnac-d)", *Journal of Near Infrared*, 14(6):403-441, 2003 (非特許文献 10)

50

F. Goge, R. Joffre, C. Jolivet, I. Ross, and L. Ranjard, " Optimization criteria in sample selection step of local regression for quantitative analysis of large soil nirs database ", Chemometrics and Intelligent Laboratory Systems, 110(1):168-176, 2012 (非特許文献 1 1)

L. Ramirez-Lopez, T. Behrens, K. Schmidt, A. Stevens, J.A.M. Dematte, and T. Scholten, " The spectrum-based learner: a new local approach for modelling soil vis-nir spectra of complex datasets ", Geoderma, 195-196:268-279, 2013 (非特許文献 1 2)

L. Ramirez-Lopez, T. Behrens, K. Schmidt, R.A. ViscarraRossel, J.A.M. Dematte, and T. Scholten, " Distance and similarity-search metrics for use with soil vis nir spectra ", Geoderma, 199:43-53, 2013 (非特許文献 1 3)

U.G. Indahl, " The geometry of pls1 explained property: 10 key noted on mathematical properties of and some alternative algorithmic approaches to pls1 modeling ", Journal of Chemometrics, 24:168-180, 2014 (非特許文献 1 4)

R.J. Pell, L.S. Ramos, and R. Manne, " The model space in partial least squares regression ", Journal of Chemometrics, 21:165-172, 2007 (非特許文献 1 5)

S. Wold, M. Hoyc, H. Martens, J. Trygg, F. Westade, J. MacGregor, and B.M. Wise, " The pls model space revisited ", Journal of Chemometrics, 23:67-68, 2009 (非特許文献 1 6)

R. Ergon, " Finding y-relevant part of x by use of pcr and pls model reduction methods ", Journal of Chemometrics, 21:537-546, 2007 (非特許文献 1 7)

R. Ergon, " Re-interpretation of nipals results solves pls inconsistency problem ", Journal of Chemometrics, 23:72-75, 2009 (非特許文献 1 8)

R.C. Martins, V.V. Lopez, P. Valentao, J.C.M.F. Carvalho, P. Isabel, M.T. Amaral, M.T. Batista, P.B. Andrade, and B.M. Silva, " Relevant principal component analysis applied to the characterization of Portuguese heather honey ", Natural Product Research, 22:1560-1582, 2007 (非特許文献 1 9)

【 0 1 2 3 】

30

40

50

【表 1】

アルゴリズム1 局所的形状及び部分空間識別

ステップA. 方向を見出す
要求: $U \in F \wedge U = xV$; ここに V は F の基底である

保証: $j = \operatorname{argmax}(X^T Y)$ 局所

```

r ← 半径           %探索半径を割り当てる
d ← 方向           %探索方向の数を割り当てる
v ← 体積           %探索方向の初期の体積を割り当てる
emax ← 誤差       %探索方向の許容可能な誤差を割り当てる
S ← 形状を計算(r, d, v, U)           %SはFの部分空間である
方向見出し ← 真
while 方向見出し = 真 do
  for S中のs do
    {e, npc, s} ← 関連PLS(Y, S)           %誤差(e)及び固有ベクトル(nPC)を最小にする
                                           vの実現再設定によって局所的に最適な学習セットを得る
    if (e < emax) かつ (npc < npcmax) then
      s ← {S|s}; E ← {E|s}; npc ← {npc|npc}           %関連する方向を記憶する
    else
      s ← 新たな探索方向(s)           %探索方向を変更する
      s ← {S|s}; E ← {E|s}; npc ← {npc|npc}           %関連する方向を記憶する
    end if
  end for
  if (max(E) < emax) かつ (max(npc) < npcmax) then
    方向見出し ← 偽           %部分空間が識別された
  end if
end while
ステップB. 部分空間の凸包の体積を最小にする
[C1, C2, C3] ← 凸包のシンプレックス(C)を生成           %3つの凸包をシンプレックス最適化用に生成する
while max(E) < emax2 do
  for C中のS do
    [e, npc, s] ← 関連するPLS(Y, S) %それぞれの凸包モデルを得る
  end for
  [C1, C2] ← Min(E)を選択 %最悪の凸包モデルを除外する
  C ← シンプレックス移動(C1, C2)
end while
出力: スペクトルxが与えられてyを予測するための部分空間S ∈ F

```

【 0 1 2 4 】

【表 2】

アルゴリズム2 特徴空間マッピング-定量化の知識ベース

要求: $U_i \in F \wedge U_i \approx x_i V$; ここに V は F の基底である

保証: $j = \operatorname{argmax}(X^T Y)$ 局所

```

r ← 半径           %探索半径を割り当てる
d ← 方向           %探索方向の数を割り当てる
v ← 体積           %探索方向の初期の体積を割り当てる
emax ← 誤差       %探索方向の許容可能な誤差を割り当てる
s ← 形状を計算(r, d, v, U)   %SはFの部分空間である
while 新たな部分空間=真 do
  while 新たな探索が可能=真 do
    アルゴリズム1を実行する
    新たな探索が可能 ← 新たな探索点()を取得
    if 新たな探索が可能 ≠ スル then
      S ← [S|s]; M ← [M|m]
    else
      新たな探索が可能 ← スル
    end if
  end while
  Unew ← 新たな座標()を選択
  if Unew == スル then
    新たな部分空間 ← 偽
  end if
end while
出力: 任意のスペクトルxによって与えられたyを予測するための  $S \in F$  の最適な部分空間マッピング

```

10

20

【 0 1 2 5 】

30

40

50

【表 3】

アルゴリズム3 特徴空間マッピング-分類の知識ベース

要求: $U_i \in F \wedge U_i = x_i V$; ここに V は F の基底である

保証: $j = \operatorname{argmax}(X^T L)_{\text{局所}} \wedge L = \text{ロジット}(P)$

```

r ← 半径           %探索半径を割り当てる
d ← 方向           %探索方向の数を割り当てる
v ← 体積           %探索方向の初期の体積を割り当てる
emax ← 誤差       %探索方向の許容可能な誤差を割り当てる
クラスタリング ← 誤差  所定のPに対するクラスタリング基準を定める
C ← 凸包を見出すs (F, L)  F全体にわたって凸包を見出す
C ← クラスタを個別にチェックする (C)  各凸包は独立でなければならない
dv ← 凸包の体積変化
for C中のc do
  while 再整形=真 do
    u ← 頂点を選択 (c)           %開始座標を探索する
    S ← 形状を計算する (r, d, v, u)  %方向及び形状を探索する
    b ← アルゴリズム1 (S)  uの部分空間の境界
    if bnew = b = ヌル then
      i ← 境界を移動 (c, d, v)     %内側へ移動する
      c ← 凸包を見出す (F, L, i)   %新たな凸包を決定する
    Else
      再整形 ← 偽
    end if
  end while
  [B|b] ← アルゴリズム2 (c, b)     %凸包境界の順次の部分空間識別を実行する
end for

```

10

20

出力: 任意のスペクトル x によって与えられる P を予測するための、 F 全体にわたる L の最適な部分空間マッピング

【 0 1 2 6 】

30

40

50

【表 4】

アルゴリズム4 局所的形状の特性化

要求: $F_1 \dots F_B \in \mathbb{F}, K_1 \dots K_B \in \mathbb{K}; f(t, u) = \arg \max_t f^t(u)$

保証: $\{w\} = \{c\} = 1, F = TW^T, K = UC^T; f = \arg \min \left\{ PRESS \times \left[\frac{R_{PC}}{R} + \frac{\beta_{sel, \text{vars}}}{R_{sel, \text{vars}}} + \frac{1}{\text{dim}(V, W)} + p\text{-value} \right] \right\}$

探索方向Dが与えられる:

while StableGroup = FALSE **do**

for D中の F_i, K_i **do**

$pc \leftarrow 1$

while pc抽出=偽 **do**

$[w, \Sigma, c] \leftarrow \text{svd}(k^T F)$

while 収束=偽 **do**

$t \leftarrow Fw / (w^T w)$

$u \leftarrow Kc / (c^T c)$

$w^* \leftarrow F^T u / (u^T u); w^* = w^* / \|w^*\|$

$c^* \leftarrow K^T t / (t^T t); c^* = c^* / \|c^*\|$

$w \leftarrow w_{i+1} \leftarrow w_i; e_c \leftarrow c_{i+1} \leftarrow c_i$

if $e_w < \text{閾値}$ **かつ** $e_c < \text{閾値}$ **then**

 収束=真

end if

end while

$p \leftarrow F^T t^* / (t^{*T} t^*); q \leftarrow K^T u^* / (u^{*T} u^*); (P|p) \leftarrow p; (Q|q) \leftarrow q$

$F \leftarrow F - tp^T; K \leftarrow K - uq^T$

$\beta_{p|ts} \leftarrow W(P^T W)^{-1} Q$

$K \leftarrow P\beta_{p|ts}$

$E_i^2 \leftarrow (K - \hat{K})^2$

if $E^2 \leftarrow \text{減少}$ **then**

$pc \leftarrow pc + 1$

else

 pc抽出=偽

end if

end while

$t_i \leftarrow T_{w, \beta_{p|ts}}$

$[F_i, K_i] \leftarrow \text{サンプルを選択}(t_i)$

end for

$[F_i, K_i] \leftarrow \text{再結合}(F, K)$

if $[F, K]$ が安定 **then**

 安定グループ=真

end if

end while

$[F, K] \leftarrow \text{変数を選択}(F)$

出力: 所定の探索方向Dに対して最適なF及びF

10

20

30

【 0 1 2 7 】

40

50

【表 5】

アルゴリズム5 直交射影を用いた部分空間情報の最適化

要求: $F_1, \dots, F_n \in \mathbb{R}^k, K_1, \dots, K_n \in \mathbb{R}^k, f(t, u) = \text{argmax}(t^t u)$

保証: $\|w\| = 1, \|e\| = 1, F = TW^t, K = UC^t, J = \text{argmin} \left(\text{PRESS} \times \left\{ \frac{npe}{n} + \frac{n_{\text{def. vars}}}{n_{\text{vars}}} + \frac{1}{\text{cov}(v, w)} + p\text{-value} \right\} \right)$

探索方向 D が与えられる

TP^t を最大にL, かつ $U_0 Q_0^t$ を最小にする f, k を、F, K から選択する

while $A < np$ do

$\{c_i, \Sigma_i, w_i\} = \text{SVD}(K^t F); w_i = w_i / \|w_i\|$

$t = Fw/w^t w; [T|t] \leftarrow t$

$p = F^t t/t^t t; [P|p] \leftarrow p$

$q = K^t t/t^t t; [Q|q] \leftarrow q$

$E_{f,i} = F_i - T_i P_i^t$

$\{Y_{o,i}, \Sigma_{o,k}, w_{o,i}\} = \text{SVD}(E_{f,i}, T_i); w_{o,i} = w_{o,i} / \|w_{o,i}\|$

$t_o = F_i w_{o,i}$

$p_o = F^t t_o/t_o^t t_o$ or $p_o = p - w_o^t p/w_o^t w_o; [P_o|p_o] \leftarrow p_o$

$F_{i+1} = F_{i+1} - T_i P_i^t - T_{o,i} P_{o,i}^t$

$E_{k,i} = K_i - T_i Q_i^t$

$\{c_{o,i}, \Sigma_{o,k}, c_{o,i}\} = \text{SVD}(E_{k,i}, T_i); c_{o,i} = c_{o,i} / \|c_{o,i}\|$

$u_o = K_i c_{o,i}$

$q = K^t u_o/u_o^t u_o; [Q_o|q_o] \leftarrow q_o$

$K_{i+1} = K_{i+1} - T_i Q_i^t - U_{o,i} Q_{o,i}^t$

end while

$F = TP^t + T_o P_o^t$

$K = TQ^t + U_o Q_o^t$

$\beta \leftarrow W(P^t W)^{-1} Q$

出力: 所定の探索方向 D に対する、最適な直交フィルタ処理した関係 $K = F \beta$

10

20

【 0 1 2 8 】

30

【表 6】

表1. 特徴空間の特性化用の測定基準

記述	式	目的
統計的表現	n	局所的方向におけるデータ数
潜在比	np/n	LVの数対局所的データの比率
複雑度の変化率	1/k	固有値の減少率
複雑度	$C = npc/(n \cdot k)$	局所的データ集合の複雑度
PRESS	$\sum(Y - \hat{Y})^2 / (n - p)$	平方和の予測値
共直線性	$\text{cov}(v, w)$	FとKの固有構造の共直線性
モデル分散	$R_F^2 = 1 - \sum E_F^2 / \sum F^2$	モデルにおけるFの分散
予測分散	$R_K^2 = 1 - \sum (TQ^t - K)^2 / \sum K^2$	モデルの分散の予測値
スペクトル干渉	$R_{F_o}^2 = \sum (T_o P_o^t)^2 / \sum F^2$	F内の直交情報の比率
定量化干渉	$R_{K_o}^2 = \sum (U_o Q_o^t)^2 / \sum K^2$	K内の直交情報の比率
共分散	$R_{F,cov}^2 = \sum (TP^t)^2 / \sum F^2$	F-K間の共分散の比率
共分散	$R_{K,cov}^2 = \sum (TQ^t)^2 / \sum K^2$	K-F間の共分散の比率
無相関情報	$1 - R_{F,cov}^2$	F内の無相関情報
無相関情報	$1 - R_{K,cov}^2$	K内の無相関情報

40

【 0 1 2 9 】

50

【表 7】

表2. 血液試料における大域的PLSのベンチマーク対自己学習方法

パラメータ	範囲	nLV	大域的PLS		自己学習方法		
			誤差 (%)	R ²	nLV ^(a)	誤差 (%)	R ²
赤血球 (10 ¹² /l)	2.0-5.5	7	11.5	0.3731	1	2.4	0.9889
ヘモグロビン (g/dl)	6.5-16.0	8	6.9	0.8395	1	1.7	0.9897
ヘマトクリット (%)	20-45.4	7	9.2	0.8392	1	3.3	0.9846
MCV (fl)	75-120	7	7.2	0.8421	1	2.8	0.9872
MCHC (g/dl)	30-37	7	8.2	0.8312	1	3.1	0.9835
白血球 (10 ⁹ /l)	3-22	10	27.0	0.4512	1	5.2	0.9907
血小板 (10 ⁹ /l)	22-753	9	6.8	0.7911	1	3.8	0.9698
ビリルビン (mg/dl)	0.25-1.5	6	12.5	0.9667	1	7.0	0.9838
グルコース (mg/dl)	60-370	10	21.0	0.7210	2	6.0	0.9865
ミオグロビン (ng/ml)	27-345	8	31.0	0.9357	1	8.0	0.9930
クレアチニン (mg/dl)	0.3-6.0	12	38.0	0.6134	3	14.0	0.9642
CRP (mg/dl)	3-14	8	12.5	0.9617	2	6.5	0.9834
トリグリセリド (mg/dl)	30-260	11	22.3	0.9156	3	8.0	0.9674
尿素 (mg/dl)	10-140	8	67.3	0.8471	2	15.0	0.9768
尿酸 (mg/dl)	1-8	7	19.0	0.7012	3	6.0	0.9479

(a) LVの数の中央値

10

20

【 0 1 3 0 】

【表 8】

表3. 血液試料における分類ベンチマーク: 真対偽の分類

下記との 関連性の確率	大域的PLS				自己学習方法			
	真 陽性 (%)	真 陰性 (%)	偽 陽性 (%)	偽 陰性 (%)	真 陽性 (%)	真 陰性 (%)	偽 陽性 (%)	偽 陰性 (%)
貧血症	45	40	4	11	53	47	0	0
白血球増加症	21	38	6	15	48	50	0	2
血栓性血小板減少性紫斑病	0	50	50	0	46	41	4	9
血小板増加症	39	50	11	0	50	50	0	0
肝不全	42	39	8	11	43	50	7	0
真性糖尿病	42	38	9	4	48	49	2	1
急性心筋梗塞	50	50	0	0	50	50	0	0
腎機能障害	37	41	13	9	43	46	8	3
炎症	50	15	0	35	50	42	0	8

(a) LVの数の中央値

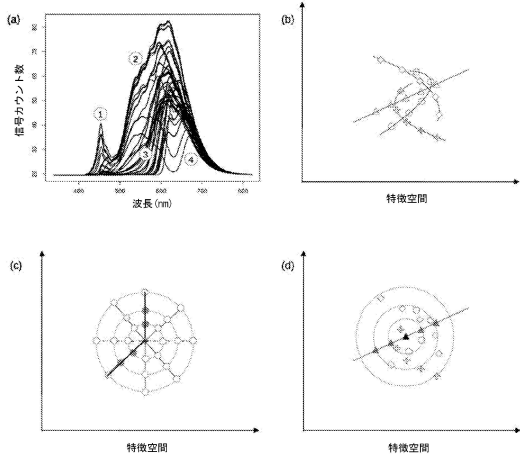
30

40

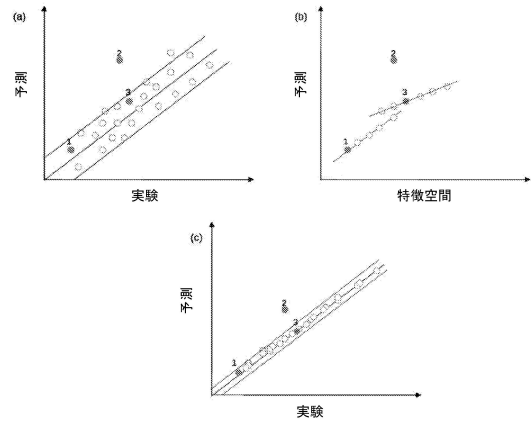
50

【図面】

【図 1】



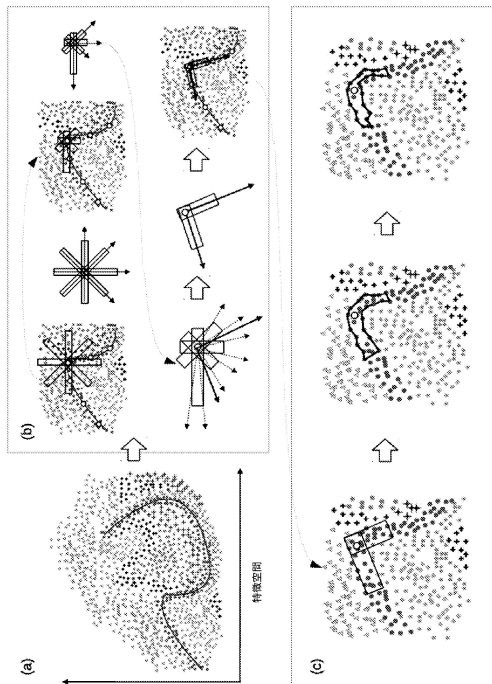
【図 2】



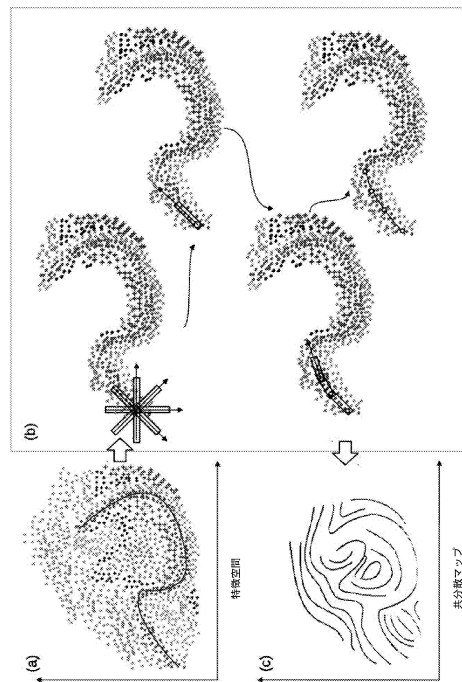
10

20

【図 3】



【図 4】

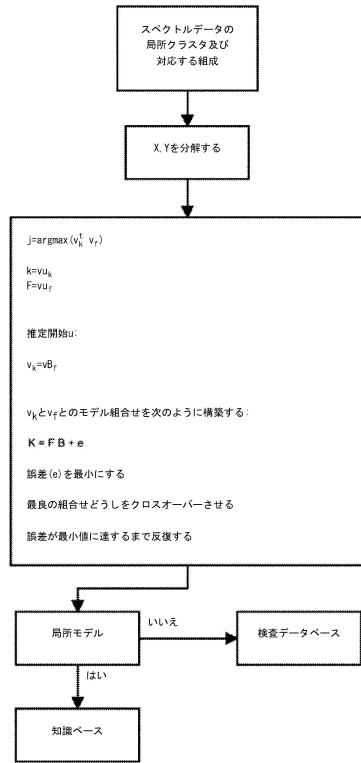


30

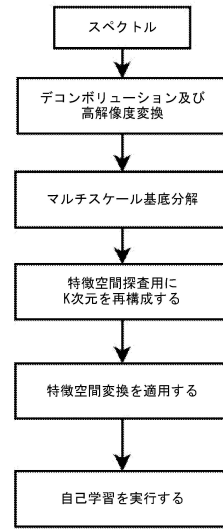
40

50

【 図 9 】



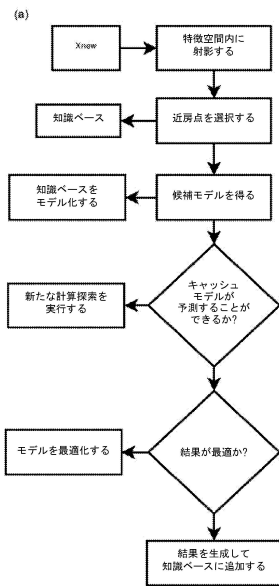
【 図 1 0 】



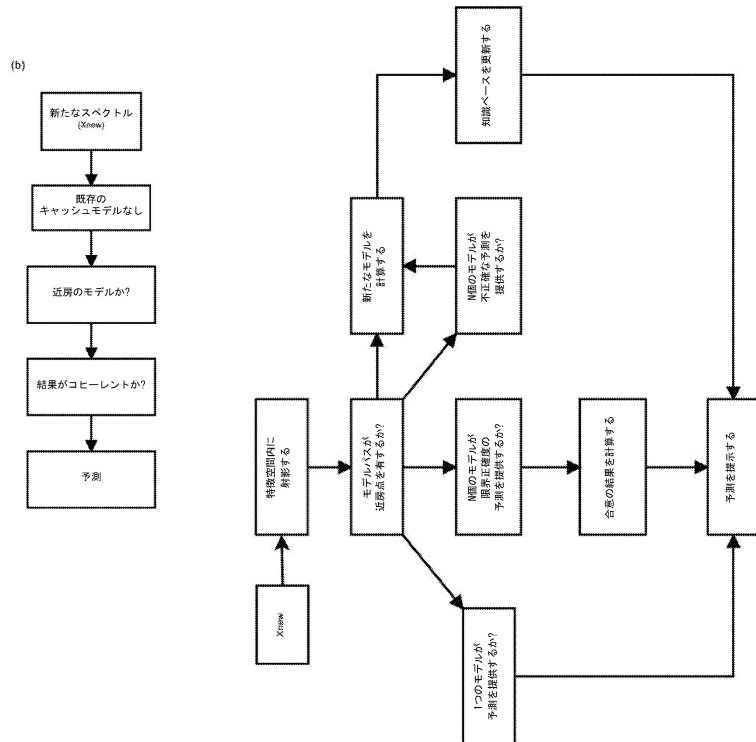
10

20

【 図 1 1 】



【 図 1 2 】

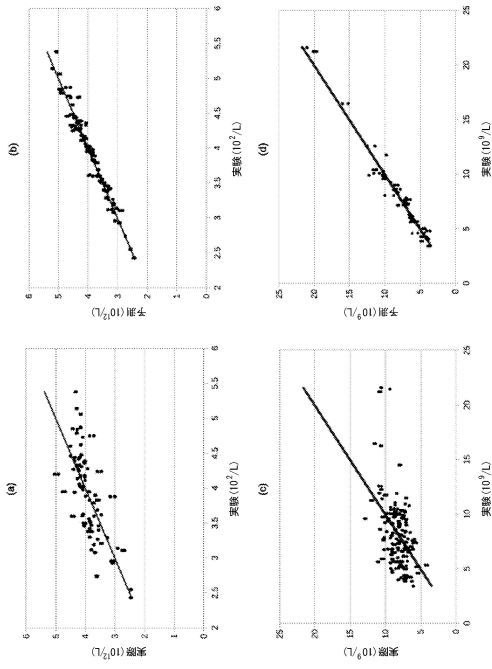


30

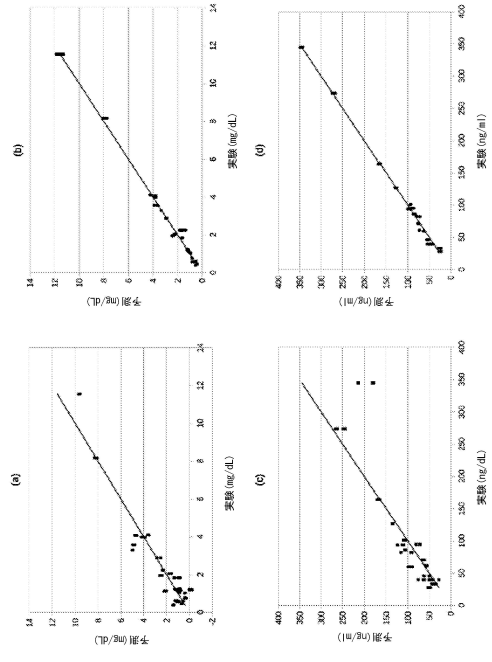
40

50

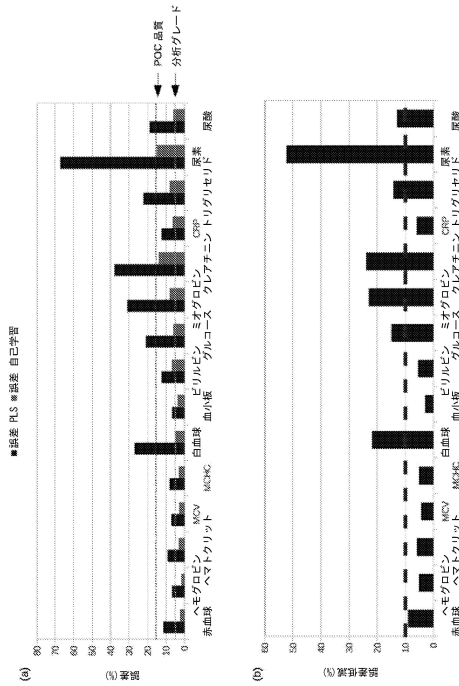
【図 1 3】



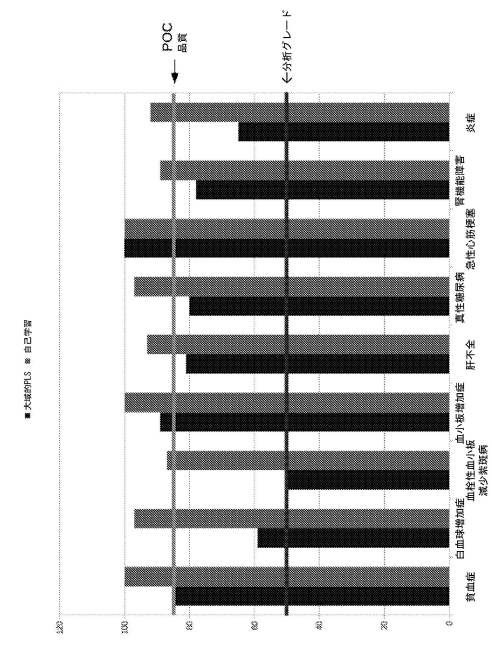
【図 1 4】



【図 1 5】



【図 1 6】



10

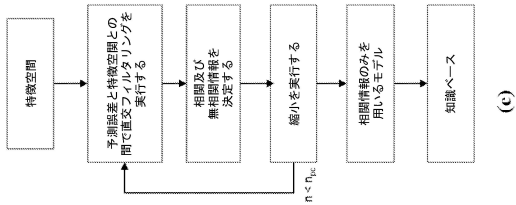
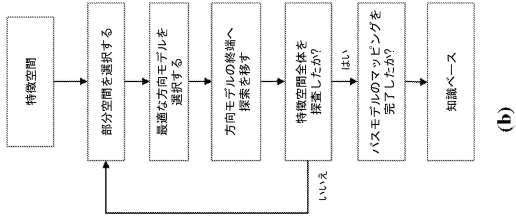
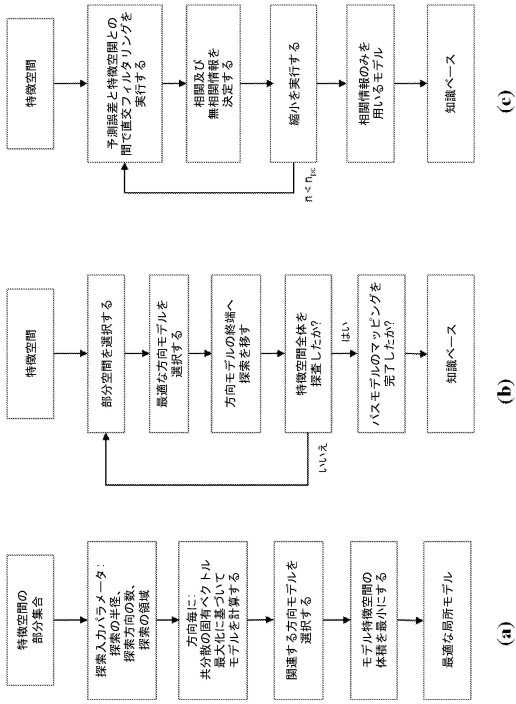
20

30

40

50

【 図 17 】



10

20

30

40

50

フロントページの続き

- 弁護士 杉村 光嗣
 (74)代理人 100169823
 弁理士 吉澤 雄郎
 (72)発明者 ルイ ミゲル ダ コスタ マルティンズ
 ポルトガル国 4 2 0 0 - 4 6 5 ポルト ルア ドトール ロベルト フリアス カンプス ダ エフイ
 ーユーピー イーエニエーエスセー テック
 審査官 小野寺 麻美子
 (56)参考文献 特開 2 0 0 3 - 0 3 5 6 6 3 (J P , A)
 特開 2 0 1 5 - 0 5 2 5 8 1 (J P , A)
 特表 2 0 0 4 - 5 2 8 0 8 3 (J P , A)
 米国特許出願公開第 2 0 0 4 / 0 0 9 8 2 0 5 (U S , A 1)
 RAMIREZ-LOPEZ,L. et al. , The spectrum-based learner: A new local approach for modeling
 soil vis-NIR spectra of complex datasets , Geoderma , 2013年 , Volumes 195-196 , pp. 26
 8-279
 (58)調査した分野 (Int.Cl. , D B 名)
 G 0 1 N 2 1 / 0 0 - G 0 1 N 2 1 / 0 1
 G 0 1 N 2 1 / 1 7 - G 0 1 N 2 1 / 6 1
 G 0 1 J 3 / 0 0 - G 0 1 J 4 / 0 4
 G 0 1 J 7 / 0 0 - G 0 1 J 9 / 0 4
 G 0 6 F 1 6 / 0 0 - G 0 6 F 1 6 / 9 5 8
 G 0 6 N 3 / 0 0 - G 0 6 N 3 / 1 2 6
 G 0 6 N 1 0 / 0 0 - G 0 6 N 9 9 / 0 0
 G 0 6 T 7 / 0 0 - G 0 6 T 7 / 9 0
 G 1 6 C 1 0 / 0 0 - G 1 6 C 9 9 / 0 0