(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2006-72666
(P2006-72666A)
(43)公開日　平成18年3月16日(2006.3.16)

(51)Int.Cl.	F I	テーマコード（参考）
G06F 17/30 (2006.01)	G06F 17/30 210D	5B075

審査請求 有　請求項の数 10　ＯＬ　（全 16 頁）

(21)出願番号　特願2004-254856 (P2004-254856)
(22)出願日　平成16年9月1日 (2004.9.1)

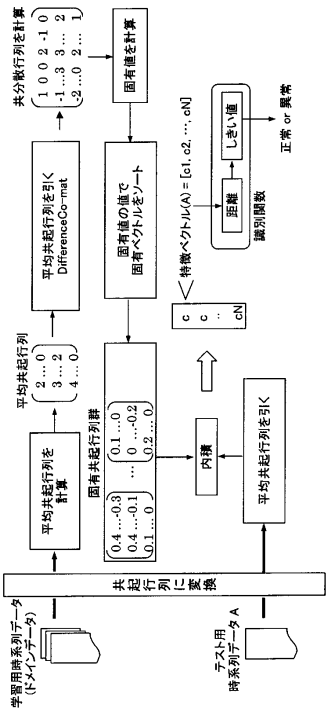
特許法第３０条第１項適用申請有り　２００４年３月１日から３日　日本ソフトウェア科学会主催の「第７回プログラミングおよび応用のシステムに関するワークショップ（ＳＰＡ２００４）」において文書をもって発表

(71)出願人　503360115
独立行政法人科学技術振興機構
埼玉県川口市本町４丁目１番８号
(74)代理人　100091443
弁理士　西浦　▲嗣▼晴
(72)発明者　岡　瑞起
茨城県つくば市天久保３－７－２　シンセイマンション２１２
(72)発明者　加藤　和彦
茨城県つくば市吾妻４－８－２
(72)発明者　大山　恵弘
東京都北区田端２－１－６　美工勤坂ハイツ２０３

Ｆターム(参考)　5B075 ND02 NR12 PQ14 PR04 PR08 QP01

(54)【発明の名称】時系列データ判定方法

(57)【要約】
【課題】　時系列データに含まれる動的情報をとらえて、時系列データが所定のカテゴリに属するものであるか否かを判定することができる時系列データ判定方法を提供する。
【解決手段】　主成分分析に用いる複数の時系列入力データを、複数種類のイベントに含まれる二種類のイベント間の関連性を共起行列で表した行列データに変換する。共起行列は、全ての２つのイベント間の関連性の強さをその距離と出現頻度により表現する。したがって本発明によれば、時系列データに含まれる動的情報を利用して、時系列データが所定のカテゴリに属するものであるか否かを従来よりも高い精度で判定することができる。
【選択図】　図１



【特許請求の範囲】
【請求項１】
複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを特徴抽出方法と識別方法とを用いて判定する時系列データ判定方法であって、
前記特徴抽出方法として、複数の時系列入力データを前記複数種類のイベントに含まれる二種類のイベント間の関連性を共起行列で表した行列データに変換したものを用いる統計的特徴抽出方法を用い、
前記識別方法として前記統計的特徴抽出方法で抽出した特徴ベクトルを識別に利用するものを用いるを実施することを特徴とする時系列データ判定方法。
【請求項２】
前記統計的特徴抽出方法が主成分分析法である請求項１に記載の時系列データ判定方法。
【請求項３】
前記複数の時系列入力データを共起行列で表した行列データに変換する際には、
前記時系列入力データをそれぞれ予め定めたデータ長さのウィンドウで切り出して複数のウィンドウ・データを取り出すウィンドウ・データ取出ステップと、
前記ウィンドウ・データから前記データ長よりも短いデータ長を有する複数のスコープ・データを時間的なずれを持って順次抽出するスコープ・データ抽出ステップと、
前記複数のウィンドウ・データを複数の前記スコープ・データに基づいて前記ウィンドウ・データに含まれる前記複数種類のイベント相互間の時系列で見た関連性の強さを示す複数の共起行列に変換する共起行列変換ステップとを実施することを特徴とする請求項１または２に記載の時系列データ判定方法。
【請求項４】
前記スコープ・データ抽出ステップでは、前記複数種類のイベントから選択した１つの種類の前記イベントが前記ウィンドウ・データに含まれる位置を基準位置として前記１つの種類のイベントに対する１以上の前記スコープ・データを抽出し、
前記共起行列変換ステップでは、前記１つの種類のイベントについての前記１以上のスコープ・データに含まれる他の１つの種類の前記イベントの数の合計値を、前記１つの種類のイベントに対する前記他の一つの種類のイベントの頻度とし、前記頻度を前記１つの種類のイベントに対する前記複数種類のイベントとの前記関連性の強さを表示する値とする変換を行って前記ウィンドウ・データを前記共起行列に変換することを特徴とする請求項３に記載の時系列データ判定方法。
【請求項５】
複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを判定する時系列データ判定方法であって、
予め学習用の複数の時系列データをそれぞれ予め定めたデータ長さのウィンドウで切り出して複数のウィンドウ・データを取り出すウィンドウ・データ取出ステップと、
前記ウィンドウ・データから前記データ長よりも短いデータ長を有する複数のスコープ・データを時間的なずれを持って順次抽出するスコープ・データ抽出ステップと、
前記複数のウィンドウ・データを複数の前記スコープ・データに基づいて前記ウィンドウ・データに含まれる前記複数種類のイベント相互間の時系列で見た関連性の強さを示す複数の共起行列に変換する共起行列変換ステップと、
前記複数の共起行列を入力として統計的特徴抽出方法により特徴ベクトルを求めるための基礎となる固有共起行列群を決定する固有共起行列群決定ステップと、
前記１以上のカテゴリを含む１以上のプロファイル学習用時系列データに対して前記ウィンドウ・データ取出ステップ、前記スコープ・データ抽出ステップ及び前記共起行列変換ステップと同様のステップをそれぞれ実施して、前記１以上のプロファイル学習用時系列データを１以上のプロファイル用共起行列に変換するプロファイル用共起行列変換ステップと、

前記１以上のプロファイル用共起行列と前記固有共起行列群とに基づいて前記１以上のプロファイル学習用時系列データについての１以上の判定用特徴ベクトルを抽出する判定用特徴ベクトル抽出ステップと、

テストの対象となるテスト時系列データに対して前記ウィンドウ・データ取出ステップ、前記スコープ・データ抽出ステップ及び前記共起行列変換ステップと同様のステップを実施して、前記テスト時系列データをテスト用共起行列に変換するテスト用共起行列変換ステップと、

前記テスト用共起行列と前記固有共起行列群とに基づいて前記テスト用時系列データについてのテスト用特徴ベクトルを抽出するテスト用特徴ベクトル抽出ステップと、

前記１以上の判定用特徴ベクトルと前記テスト用特徴ベクトルとに基づいて、前記テスト時系列データが前記１以上のカテゴリに属するか否かを判定する判定ステップとからなる時系列データ判定方法。

【請求項６】

前記スコープ・データ抽出ステップでは、前記複数種類のイベントから選択した１つの種類の前記イベントが前記ウィンドウ・データに含まれる位置を基準位置として前記１つの種類のイベントに対する１以上の前記スコープ・データを抽出し、

前記共起行列変換ステップでは、前記１つの種類のイベントについての前記１以上のスコープ・データに含まれる他の１つの種類の前記イベントの数の合計値を、前記１つの種類のイベントに対する前記他の一つの種類のイベントの頻度とし、前記頻度を前記１つの種類のイベントに対する前記他の種類のイベントの前記関連性の強さを表示する値とする変換を行って前記ウィンドウ・データを前記共起行列に変換することを特徴とする請求項５に記載の時系列データ判定方法。

【請求項７】

前記判定用特徴ベクトル抽出ステップでは、前記プロファイル用共起行列と前記固有共起行列群とをベクトルかした後にその内積を求めて前記判定用特徴ベクトルを決定し、

前記テスト用特徴ベクトル抽出ステップでは、前記テスト用共起行列と前記固有共起行列群とをベクトル化した後にその内積を求めて前記テスト用特徴ベクトルを抽出することを特徴とする請求項５に記載の時系列データ判定方法。

【請求項８】

前記判定ステップでは、所定のベクトル識別関数を用いて前記テスト用時系列データと前記判定用特徴ベクトルとのユークリッド距離が閾値以内であるか否かにより前記テスト時系列データが前記１以上のカテゴリに属するか否かを判定する請求項５に記載の時系列データ判定方法。

【請求項９】

前記学習用の複数の時系列データに、前記テスト時系列データを含めて、前記固有共起行列群を更新することを特徴とする請求項５に記載の時系列データ判定方法

【請求項１０】

請求項１乃至９のいずれか１項に記載の時系列データ判定方法を用いて、コンピュータシステムに入力される時系列データの異常を判別することを特徴とする時系列データ異常判別方法。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、時系列データが所定の１以上のカテゴリに属するものであるか否かを判定する時系列データ判定方法に関するものである。

【背景技術】

【０００２】

ユーザのパスワードを盗み出し、そのユーザになりすまして不正にコンピュータを使用する、いわゆる「なりすまし」を検出するためには、コンピュータに入力される時系列データに異常があるか否か（入力された時系列データが、なりすまし者によって作成された

時系列データであるか否か）を異常検知システムで検知することが効果的である。公知の異常検知システムでは、最初にユーザの典型的な行動を示すプロファイル（ユーザが作成した時系列データに現れる特徴）を作成する。そしてテスト対象である入力データ（時系列データ）のプロファイルをそのユーザのプロファイルと比較することにより、正常なユーザが作成した時系列データであるか、なりすまし者が作成した異常な時系列データであるかを識別する。

【０００３】

典型的な検査対象となる入力データは、使用されたUNIX（登録商標）コマンド、アクセスされたファイル等の時系列データ等である。入力された時系列データが、正常か異常かを識別する過程は２つのステップに分けられる。まず第１ステップにおいて、時系列データの特徴抽出を行う。そして第２ステップにおいて、抽出された特徴が正常か異常かを識別する。

【０００４】

第１のステップの特徴抽出を行う代表的な従来手法には、ヒストグラム（Histogram）とエヌグラム（N-gram）とがある。ヒストグラム（Histogram）では、時系列データに現れる項目（イベント）の出現頻度ベクトルが抽出の対象となる特徴ベクトルとなる。また、エヌグラム（N-gram）では、連続するN個の項目を１つの特徴とする［非特許文献１乃至３］。

【０００５】

また第２ステップの抽出された特徴を正常か異常かを識別する手法としては、これまでに様々な手法が提案されている。それらの中で代表的な手法には、ルールベース［非特許文献４］、オートマトン［非特許文献５］、ベイジアンネットワーク［非特許文献６］、Naiveベーズ［非特許文献７］、ニューラルネットワーク［非特許文献８］、マルコフモデル［非特許文献９］、隠れマルコフモデル［非特許文献１０］とがある。

【非特許文献１】Ｙｅ，Ｘ．Ｌｉ，Ｑ．Ｃｈｅｎ，Ｓ．Ｍ．Ｅｍｒａｎ，及びＭ．Ｘｕ著の「Ｐｒｏｂａｂｌｉｓｔｉｃ Ｔｅｃｈｎｉｑｕｅｓ ｆｏｒ Ｉｎｔｒｕｓｉｏｎ Ｄｅｔｅｃｔｉｏｎ Ｂａｓｅｄ ｏｎ Ｃｏｍｐｕｔｅｒ Ａｕｄｉｔ Ｄａｔａ」ＩＥＥＥ Ｔｒａｎｓａｃｔｉｏｎｓ ｏｆ Ｓｙｓｔｅｍｓ Ｍａｎ ａｎｄ Ｃｙｂｅｒｎｅｔｉｃｓ，Ｖｏｌ．３１，ｐｐ．２６６－２７４，２００１

【非特許文献２】Ｓ．Ａ．Ｈｏｆｍｅｙｒ，Ｓ．Ｆｏｒｒｅｓｔ及びＡ．Ｓｏｍａｙａｊｉ著の「Ｉｎｔｒｕｓｉｏｎ Ｄｅｔｅｃｔｉｏｎ ｕｓｉｎｇ Ｓｅｑｕｅｎｃｅｓ ｏｆ Ｓｙｓｔｅｍ Ｃａｌｌｓ」Ｊｏｕｒｎａｌ ｏｆ Ｃｏｍｐｕｔｅｒ Ｓｅｃｕｒｉｔｙ，ｖｏｌ．６，ｐｐ．１５１－１８０，１９９８

【非特許文献３】Ｗ．Ｌｅｅ及びＳ．Ｊ．Ｓｔｏｌｆｏ著の「Ａ ｆｒａｍｅｗｏｒｋ ｆｏｒ ｃｏｎｓｔｒｕｃｔｉｎｇ ｆｅａｔｕｒｅｓ ａｎｄ ｍｏｄｅｌｓ ｆｏｒ ｉｎｔｒｕｓｉｏｎ ｄｅｔｅｃｔｉｏｎ ｓｙｓｔｅｍｓ」，Ｉｎｆｏｒｍａｔｉｏｎ ａｎｄ Ｓｙｓｔｅｍ Ｓｅｃｕｒｉｔｙ，ｖｏｌ．３，ｐｐ．２２７－２６１，２０００

【非特許文献４】N.Habra,B.L.Charlier,A.Mounji及びI.Mathieu著の「ASAX:Software Architecture and Rule-BasedL anguage for Universal Audit Trail Analysis」In Proc.of European Symposiumon Researchin Computer Secu-rity(ESORICS),pp.435-450,1992

【非特許文献５】R.Sekar,M.Bendre及びP.Bollineni著の「A Fast Automaton Based Method for DetectingAnomalous Program Behaviors」In Proceedings of the 2001 IEEE Symposium on Security and Privacy,pp.144-155,Oakland,May2001.

【非特許文献６】W.DuMouchel著の「Computer Intrusion Detection Based on Bayes Factors for Comparing Command Transition Probabilities」Technical Report TR91,National Institute of Statistical Sciences(NISS),1999.

【非特許文献７】R.A.Maxion 及びT.N.Townsend.著の「Masquerade Detection Using Truncated CommandLines」In Prof.of the International Conferenceon Dependable Systems and Networks(DSN-02),pp.219-228,2002.

【非特許文献８】A.K.Ghosh ,A.Schwartzbard,及びM.Schatz著の「A study in using neural networks foranomaly and misuse detection」In Proc.of USENIX Security Symposium,pp.141-151,1999.
【非特許文献９】J.S.Tan,K.M.C.及びR.A.Maxion.MarkovChains著の「Classifiers and Intrusion Detection.In Proc.of 14th IEEE Computer Security Foundations Workshop,pp.206-219,2001
【非特許文献１０】C.Warrender,S.Forresto及びB.A.Pearlmutter著の「Detecting Intrusions using System Calls :Alternative Data Models」In IEEE Symposium on Security and Privacy,pp.133-145,1999.
【発明の開示】
【発明が解決しようとする課題】
【０００６】
しかしながらヒストグラム（Histogram）では、時系列データに現れる項目（イベント）の出現頻度ベクトルが特徴となる。また、エヌグラム（N-gram）では、連続するN個の項目を１つの特徴とする。しかしながらこれらの従来手法では、時系列データにおけるユーザの挙動の動的情報（時系列で見たユーザの挙動に関する情報即ちイベント時系列上に現れるイベントの種類とそれらの出現順で定まる各ユーザの特異的な特徴）が利用できない又は時系列データにおけるユーザの挙動の動的情報が失われるという問題や、単独もしくは隣接するイベントの特徴しか利用できない又は隣接するイベント間の特徴しか表現されないという問題がある。
【０００７】
本発明の目的は、時系列データに含まれる動的情報をとらえて、時系列データが所定のカテゴリ（特徴）を含むものであるか否かを判定することができる時系列データ判定方法を提供することにある。
【０００８】
本発明の他の目的は、従来よりも判定精度の高い時系列データ判定方法を提供することにある。
【０００９】
本発明の別の目的は、時系列データに異常があるか否かを判別することができる時系列データ異常判別方法を提供することにある。
【課題を解決するための手段】
【００１０】
本発明は、Eigen Co-occurrence Matrix（ECM）手法を開発したことを基礎としてなされたものである。このECM手法は、まず時系列情報を考慮しながら、時系列データに含まれるイベント間の関連付けを行う。この関連付けは、二つのイベント間の関連に着目し、全ての二項間イベントの関連性をCo-occurrence Matrix(共起行列)として表現することにより行う。共起行列は時系列データに現れる項目（イベント）間全ての関係性が表現することができる。これは、ヒストグラム（Histogram）やエヌグラム（N-gram）では表現することができなかった時系列データの特徴である。具体的な発明では、共起行列に対し主成分分析を行い、直交する主成分ベクトル空間を生成する。それぞれの共起行列は、主成分ベクトル空間上のベクトルとして特徴が抽出される。特徴をベクトルとして抽出することにより、様々なベクトル識別関数を利用することも可能になる。
【００１１】
本発明の時系列データ判定方法は、複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを特徴抽出方法と識別方法とを用いて判定する。本発明では、特に、前記特徴抽出方法として、複数の時系列入力データを複数種類のイベントに含まれる二種類のイベント間の関連性を共起行列で表した行列データに変換したものを用いる統計的特徴抽出方法を用いる。そして識別方法として統計的特徴抽出方法で抽出した特徴ベクトルを識別に利用するものを用いる。ここで複数種類のイベントとは、時系列データを構成する複数の項目を意味し、時系列データが複数のコマン

ドから構成されている場合には、その複数のコマンドがそれぞれイベントである。またカテゴリとは、上位概念で見れば時系列データの種別を意味する概念であって、下位の概念で見れば時系列データから得られる後述する特徴ベクトルの集合が属する種別である。例えば、ある時系列データがある正常であるか否かは、時系列データが予め定めた１以上のカテゴリに属するか否かにより判断することができる。なお特徴ベクトルとカテゴリとの関係で見れば、特徴ベクトルが存在する空間の部分領域に対応するものがカテゴリとなる。統計的特徴抽出方法としては、特徴ベクトルを抽出できるものであれば、どのようなものでよく、例えば主成分分析法を用いることができる。また特徴ベクトルを利用して時系列データがどのカテゴリに属するのかを判定する識別方法は任意である。従来技術の欄に記載した公知の各種の識別方法を用いることができるのは勿論である。

【００１２】

本発明の方法で採用する共起行列は、時系列データに現れる項目（イベント）間全ての関係性を表現することができる。言い替えると、共起行列は、全ての二項間の関連性の強さをその距離と出現頻度により表現する。したがって本発明によれば、時系列データに含まれる動的情報を利用して、時系列データが所定のカテゴリに属するか否かを従来よりも高い精度で判定することができる。

【００１３】

複数の時系列入力データを共起行列で表した行列データに変換する際には、ウィンドウ・データ取出ステップと、スコープ・データ抽出ステップと、共起行列変換ステップとを実施する。ウィンドウ・データ取出ステップでは、時系列入力データをそれぞれ予め定めたデータ長さのウィンドウで切り出して複数のウィンドウ・データを取り出す。ウィンドウのデータ長さは、時系列データの長さに応じて定めればよい。スコープ・データ抽出ステップでは、ウィンドウ・データからウィンドウ・データのデータ長よりも短いデータ長を有する複数のスコープ・データをデータ列上において時間的なずれを持って順次抽出する。具体的なスコープ・データ抽出ステップでは、複数種類のイベントから選択した１つの種類のイベントがウィンドウ・データに含まれる位置を基準位置として１つの種類のイベントに対する１以上のスコープ・データを抽出することができる。また共起行列変換ステップでは、複数のウィンドウ・データを複数のスコープ・データに基づいてウィンドウ・データに含まれる複数種類のイベント相互間の時系列で見た関連性の強さを示す複数の共起行列に変換する。具体的な、共起行列変換ステップでは、１つの種類のイベントについての１以上のスコープ・データに含まれるその１つの種類のイベントまたは他の種類のイベントの数の合計値を、１つの種類のイベントに対する一つの種類のイベントの頻度とし、この頻度を１つの種類のイベントに対する一つの種類のイベントの関連性の強さを表示する値とする変換を行ってウィンドウ・データを共起行列に変換する。このようにして共起行列の変換を行うと、時系列で見たイベント相互間の関連性をより適格に示す共起行列を得ることができる。

【００１４】

正当なユーザとなりすまし者を本発明の方法を利用して識別するには、さらに共起行列をパターンとして扱い、統計的パターン認識手法（識別方法）を適用することが妥当である。最も簡単なパターン認識手法（識別方法）は、パターン間のマッチングに基づく手法である。しかし共起行列そのものをパターンとして扱った場合、パターンの次元が膨大になってしまう。そのため、パターン間のマッチングでは、特徴を抽出し（情報圧縮にもなっている）、認識を行うことがより有効である。パターンから有効な特徴抽出を行うことにより、入力パターンの変動に対して頑健な認識結果が期待できる。そこで本発明のより具体的な方法では、特徴抽出方法として、主成分分析を用いて、共起行列からの特徴ベクトルの抽出に利用する。主成分分析はベクトル形式のデータを少数の特徴（主成分）で表すことを可能とする統計的特徴抽出方法である。なお主成分分析を用いた認識の成功例として、Ｔｕｒｋ等［Ｍ．Ｔｕｒｋ，Ａ．Ｐｅｎｔｌａｎｄ，「Ｅｉｇｅｎｆａｃｅｓ ｆｏｒ Ｒｅｃｏｇｕｎｉｔｉｏｎ」Ｊｏｕｒｎａｌ ｏｆ Ｃｏｇｎｉｔｉｖｅ Ｎｅｕｒｏｓｃｉｅｎｃｅ，ｖｏｌ３，Ｎｏ．１，１９９１］が提案したＥｉｇｅｎｆａｃｅ（

固有顔）による顔画像の認識が広く知られている。本発明の具体的方法では、共起行列（Ｃｏ－ｏｃｃｕｒｒｅｎｃｅ　Ｍａｔｒｉｘ）を顔画像と見なしたところにユニークな着眼点がある。

【００１５】

そこで複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを判定する本発明の具体的な時系列データ判定方法では、前述のウィンドウ・データ取出ステップと、前述のスコープ・データ抽出ステップと、前述の共起行列変換ステップに加えて、更に固有共起行列群決定ステップと、プロファイル用共起行列変換ステップと、判定用特徴ベクトル抽出ステップと、テスト用共起行列変換ステップと、テスト用特徴ベクトル抽出ステップと、判定ステップとを用いる。

【００１６】

固有共起行列群決定ステップでは、複数の共起行列を入力として主成分分析により特徴ベクトルを求めるための基礎となる固有共起行列群を決定する。またプロファイル用共起行列変換ステップでは、１以上のカテゴリを含む１以上のプロファイル学習用時系列データに対してウィンドウ・データ取出ステップ、スコープ・データ抽出ステップ及び共起行列変換ステップと同様のステップをそれぞれ実施して、１以上のプロファイル学習用時系列データを１以上のプロファイル用共起行列に変換する。また判定用特徴ベクトル抽出ステップでは、１以上のプロファイル用共起行列と固有共起行列群とに基づいて１以上のプロファイル学習用時系列データについての１以上の判定用特徴ベクトルを抽出する。更にテスト用共起行列変換ステップでは、テストの対象となるテスト時系列データに対してウィンドウ・データ取出ステップ、スコープ・データ抽出ステップ及び共起行列変換ステップと同様のステップを実施して、テスト時系列データをテスト用共起行列に変換する。またテスト用特徴ベクトル抽出ステップは、テスト用共起行列と固有共起行列群とに基づいてテスト用時系列データについてのテスト用特徴ベクトルを抽出する。そして判定ステップでは、１以上の判定用特徴ベクトルとテスト用特徴ベクトルとに基づいて、テスト時系列データが１以上のカテゴリを含むか否かを判定する。本発明の具体的な方法のように、主成分分析を介することにより、固有顔に対応する固有共起行列群（Ｅｉｇｅｎ　Ｃｏ－ｏｃｃｕｒｒｅｎｃｅ　Ｍａｔｒｉｘ）を作成すると、もとの共起行列を低次元で近似して表現することが可能になった。

【００１７】

なお判定ステップでは、具体的には、所定のベクトル識別関数を用いてテスト用時系列データと判定用特徴ベクトルとのユークリッド距離が閾値以内であるか否かによりテスト時系列データが１以上のカテゴリを含むか否かを判定する。このようなベクトル識別関数を用いると、簡単により高い精度で判定を行える。

【００１８】

精度の良い異常検知システムを構築するためには、ユーザのプロファイルを、コンセンプチュアル・ドゥリフト（Conceptual Drift）に対応させて更新する必要がある。従来の方法においては、ユーザのプロファイルを更新する際は、識別関数における結果を利用して行う必要がある(フィードバック更新)。そのため、識別関数の結果が間違っていた場合、プロファイルが正しく更新されないという問題がある。そこで本発明では、学習用の複数の時系列データに、テスト時系列データを含めて、固有共起行列群を更新すると、識別関数の結果を利用せずプロファイルの更新が可能である（フィードフォーワード更新）。したがって更新を確実に行うことができる。

【００１９】

また本発明の時系列データ判定方法を用いて、コンピュータシステムに入力される時系列データの異常を判別すると、従来よりも高い精度で異常な時系列データを判別することができる。

【発明の効果】

【００２０】

本発明によれば、時系列データに含まれる動的情報を利用して、時系列データが所定の

カテゴリを含むものであるか否かを従来よりも高い精度で判定することができる。

【発明を実施するための最良の形態】

【００２１】

以下図面を参照して本発明の実施の形態を詳細に説明する。図１は、複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを主成分分析法を用いて判定する本発明の時系列判定方法の実施の形態の一例を実施するためのプログラムの構成を示す図である。本実施の形態では、特徴ベクトルを得るために用いる固有共起行列群を得るための学習用の複数の時系列データと、プロファイル学習用の時系列データ（以下プロファイル学習用時系列データと言う）と、テストの対象となるテスト時系列データ（以下テスト用時系列データと言う）を共起行列に変換する。ここで共起行列とは、時系列データを構成する複数種類のイベントに含まれる二種類のイベント間の関連性を行列データに変換したものである。

【００２２】

時系列データを共起行列に変換するステップについて説明する。図２は、複数の学習用時系列データ［この場合にはユーザ（コンピュータにアクセスして時系列データを送信してくる人または他のコンピュータ）１乃至ユーザ３からそれぞれ送られた３つの時系列データ］の構成の一例を示している。この例では、各ユーザからの時系列データは、２０のコマンド（イベント）によってそれぞれ構成されている。後に説明するように、この実施の形態では、２０のコマンドからなる時系列データを１０のコマンド（データ長）を有するウィンドウで区切る（ウィンドウ・データ取出ステップ）。このウィンドウ・データ取出ステップでは、各時系列入力データをそれぞれ予め定めたデータ長（１０個のコマンド分のデータ長）のウィンドウで切り出して２つのウィンドウ・データを取り出す。なおウィンドウのデータ長さは、時系列データの長さに応じて定めればよい。

【００２３】

次に、ある区間の時系列データに現れる２つのイベント間の因果関係を表すために共起行列に変換する。共起行列のそれぞれの要素は、２つのイベント間の因果関係の強さを表すものである。共起行列を作成するために、ウィンドウサイズｗ、スコープサイズｓ、そしてイベントセットＢ＝｛ｂ１、ｂ２、ｂ３、・・，ｂｍ｝を定義する．ここでｍは、イベント数を示す。ウィンドウサイズｗは、１つの特徴ベクトルを抽出するイベント時系列のサイズを決定し、スコープサイズｓは、２つのイベントの因果関係を考慮する間隔幅を決定する。図２に示すデータ例では、ｗを１０、ｓを６と定義した．また、Ｂは、3人全ての学習用の時系列データ（ドメインデータ）に現れるユニークな８つのコマンド（イベント）（ｍ＝８）とする。８つのコマンドは、cd, ls, less, emacs, gcc, gdb, mkdir, cpである。２つのイベント間の因果関係または関連性の強さは、イベント間の距離と、それらが現れる頻度により定義される。つまり、注目するイベントが、ウィンドウサイズ（１０）の中で、スコープサイズ（６）以内に現れる頻度を数えることにより、イベント間の因果関係の強さを定義する。図２の例では、それぞれにユーザ一人について２つの共起行列が作られることになる。図３のウィンドウ１におけるイベントcd とイベントls の要素または頻度数７は、ウィンドウサイズ（１０）で、スコープサイズ（６）以内に、lsがcdの後に7回現れたことを示している。イベントペア（cd ls）と{ls cd}が図3のウィンドウ１において最も大きな要素または頻度数を持つ。これはこの時系列において、これらのイベントは強い関係性があることを示している．共起行列は、時系列データに現れる全ての２つのイベント相互間の因果関係または関連性の強さを表現することになる。

【００２４】

図３について、本発明との関係で、詳しく説明する。まず各ユーザの時系列データ毎に、図３に示すように、前述のウィンドウ・データから複数のスコープ・データを抽出する（スコープ・データ抽出ステップ）。このステップでは、ウィンドウ・データからウィンドウ・データのデータ長よりも短いデータ長を有する複数のスコープ・データをデータ上における時間的なずれを持って順次抽出する。この例では、６個のコマンド分のデータ長を有するスコープ・データを順次抽出している。具体的には、ウィンドウ・データを構成

する１０個のコマンドに含まれる複数種類のイベント（図３の場合には、ｃｄ，ｌｓ，ｌｅｓｓ）から選択した１つの種類のイベント（例えばｃｄ）が、ウィンドウ・データに含まれる位置を基準位置として１つの種類のイベントに対する１以上のスコープ・データを抽出する。図３の例で見れば、イベントｃｄに着目した場合、ウィンドウ１の先頭にあるイベントｃｄを含まずにこのｃｄ（基準位置）より後の６個のコマンド（イベント）を第１のスコープ・データとして抽出し、次に先頭から６番目にあるイベントｃｄを含まずにこのｃｄ（基準位置）より後の６個のコマンド（イベント）を第２のスコープ・データとして抽出する。なお図３の例のように、ウィンドウ１内に１０個しかイベントが無い場合、第２のスコープ・データでは４個のイベントを抽出する。同様に、先頭から８番目及び第９番目のイベントｃｄを基準位置にして第３及び第４のスコープ・データを抽出する。

【００２５】

次に、ウィンドウ・データから抽出した複数のスコープ・データに基づいてそのウィンドウ・データに含まれる複数種類のイベント相互間の時系列で見た関連性の強さ（二つのイベントの相互間の関連強さ）を、関連性を見る二つのイベントが現れる頻度と距離として表現する。例えば、１つの種類のイベントｃｄについての１以上（図３の場合には４つ）のスコープ・データに含まれる１つの種類のイベント（図３の場合には同じ種類のｃｄ）の数の合計値を、１つの種類のイベントに対する一つの種類のイベントの頻度とする。そして、この頻度を１つの種類のイベントに対する一つの種類のイベントの関連性の強さを表示する値とする変換を行ってウィンドウ・データを共起行列に変換する。図３の例において、ウィンドウ１中のイベントｃｄとイベントｃｄとの間の関連性を頻度として見る。前述の第１のスコープ・データ中には、１つのｃｄが含まれており、第２のスコープ・データ中には２つのｃｄが含まれており、第３のスコープ・データ中には１つのｃｄが含まれており、第４のスコープ・データ中にはｃｄはふくまれない。したがってイベントｃｄに対するイベントｃｄの頻度は、１＋２＋１＋０＝４と計算できる。同様にしてイベントｃｄに対するイベントｌｓの関連性ついてみれば、前述の第１のスコープ・データ中には、３つのｌｓが含まれており、第２のスコープ・データ中には２つのｌｓが含まれており、第３のスコープ・データ中には１つのｌｓが含まれており、第４のスコープ・データ中には１つのｌｓが含まれている。したがってイベントｃｄに対するイベントｌｓの頻度は、３＋２＋１＋１＝７と計算できる。これらの頻度には、スコープ・データを設定することにより、時間または距離の関係即ち時系列データに含まれる動的情報が含まれることになる。図３の右側領域には、ウィンドウ１及び２をそれぞれ共起行列に変換した行列データが示されている。このように時系列データを共起行列で表現すると、人間の流動的な行動のモデル化が可能になる。

【００２６】

正当なユーザとなりすまし者を本発明の方法を利用して識別するには、共起行列をパターンとして扱い、統計的特徴抽出方法として主成分分析を用いて特徴ベクトルを求め、その後特徴ベクトルを識別に利用して識別を実行する。主成分分析はベクトル形式のデータを少数の特徴（主成分）で表すことを可能とする統計的特徴抽出方法であり、主成分分析とは多変量で表されるデータの統計から、一次結合で表現される新たな変量を構成し、互いに無相関な「主成分」に要約する手法である。本実施の形態では、共起行列を先に述べたＴｕｒｋ等が提案したＥｉｇｅｎｆａｃｅ（固有顔）による顔画像と見なしている。そこで本出願においては、本発明の時系列データ判定方法をEigen Co-occurrence Matrix（ECM）手法と呼ぶ。

【００２７】

図１に示すように、時系列データから、固有共起行列群を作成する学習用の時系列データを選びこれをドメインデータとする。１つのウィンドウから変換した共起行列を前述のＭ．Ｔｕｒｋ等が発表したＥｉｇｅｎｆａｃｅ（固有顔）における顔画像と見なし、Ｅｉｇｅｎｆａｃｅに対応するＥｉｇｅｎ Ｃｏ－ｏｃｃｕｒｒｅｎｃｅ Ｍａｔｒｉｘ（固有共起行列）を作成する。主成分分析により、固有値とそれに対応する固有ベクトルが得られる。そして固有値を降順に並べ、それと対応する固有ベクトルを上からＮ個選択し、

行列化し固有共起行列群とする。

【００２８】

共起行列からの主成分分析を用いた特徴ベクトル抽出は次に述べる手順で行う。まず学習用の時系列データから得たｐ枚の学習用の共起行列のうちｉ番目の共起行列を、各要素の値を並べたＮ次元のベクトルｘｉとして表現する。ここでｐはサンプル数であり、Ｎはイベント数の２乗である。ｐ枚の共起行列の平均ベクトルを平均共起行列として下記の式で求める。ここで平均共起行列は、イベントペア（2項間）の関係性を示す。

【数１】

$$\bar{x} = \frac{1}{p}\sum_{i=1}^{p} x_i$$

【００２９】

そして各共起行列から平均共起行列（平均ベクトル）を引いたベクトルを

【数２】

$$\tilde{x}_i = x_i - \bar{\bar{x}}$$

【００３０】

で表す。この平均共起行列を引く意味は、座標軸を原点に設定するためである。そして各共起行列から平均共起行列（ｍ×ｍ行列）を引き、ベクトル化した（ｍ×ｍの行列をm^2次元の縦ベクトルにする）共起行列の集合を行列

【数３】

$$\tilde{X} = [\tilde{x}_1, \cdots, \tilde{x}_p]$$

【００３１】

で表す。この行列とその転置行列をかけた行列が図１における共分散行列($m^2 \times m^2$行列)である。

【００３２】

次に、学習用の共起行列の集合を最適に近似する正規直交基底ａを、［数３］で表した行列Ｘの共分散行列の固有ベクトルで構成する。そのために共分散行列から固有値及び固有ベクトルを計算する（$m^2 \times m^2$行列の固有ベクトルを計算）する。ここで固有値は、特徴の強さを表す。また固有ベクトルは、お互いに無相関な特徴の軸を表している。このとき、ａの各固有ベクトルａｌを、固有共起行列（Ｅｉｇｅｎ　ｃｏ－ｏｃｃｕｒｒｅｎｃｅ　ｍａｔｒｉｘ）とし、その集合を固有共起行列群（主成分）と言う。

【００３３】

具体的には、固有値を降順にソートし、それらに対応する固有ベクトルを得る（m^2個の固有ベクトルのうちＮ個のみ選択する。固有値によって、固有ベクトルをソートすることにより、特徴の強い軸を上から順番に取り出すことができる。Ｎ個の固有ベクトルをそれぞれ行列化し(m^2次元のベクトルをｍ×ｍの行列にする)、これを固有共起行列群とする。ここである共起行列ｘに対する特徴ベクトル（Ａ）（または主成分スコアＣ）を縦ベクトル化した共起行列ｘと正規直交基底ａの内積を計算することにより求める。特徴ベクトルの各成分c_1，c_2，．．．，c_Nは、共起行列ｘを表現するための各固有共起行列の貢献度を表すことになる。本実施の形態のように、特徴ベクトルを共起行列から抽出した場合、様々なベクトル空間手法を用いた特徴ベクトルの識別に使用することができる.

本発明の時系列データの判定方法と関係する部分について以下に説明する。判定方法では、前述の共起行列の変換で用いたウィンドウ・データ取出ステップと、前述のスコープ・データ抽出ステップと、前述の共起行列変換ステップに加えて、更に固有共起行列決定

ステップと、プロファイル用共起行列変換ステップと、判定用特徴ベクトル抽出ステップと、テスト用共起行列変換ステップと、テスト用特徴ベクトル抽出ステップと、判定ステップとを実施する。

【００３４】

まず固有共起行列決定ステップでは、前述のようにして複数の共起行列（学習用の時系列データを共起行列に変換したもの）を入力として主成分分析により特徴ベクトルを求めるための基礎となる固有共起行列群（固有共起行列の集合即ち主成分）を決定する。

【００３５】

そしてプロファイル用共起行列変換ステップでは、１以上のカテゴリを含む１以上のプロファイル学習用時系列データに対して先に説明したのと同様のウィンドウ・データ取出ステップ、スコープ・データ抽出ステップ及び共起行列変換ステップと同様のステップをそれぞれ実施して、１以上のプロファイル学習用時系列データを１以上のプロファイル用共起行列に変換する。ここでプロファイル学習用時系列データとしては、正常なユーザが作成したものであることが明確に判っている時系列データを用いる。学習用の時系列データからこのプロファイル学習用時系列データを選んでもよいのは勿論である。あるコンピュータにアクセスするユーザが１００人いれば、その１００人が作成した時系列データをプロファイル学習用時系列データとしてそれぞれプロファイル用共起行列に変換する。

【００３６】

次に判定用特徴ベクトル抽出ステップでは、プロファイル用共起行列と固有共起行列群とに基づいて各プロファイル学習用時系列データについての判定用特徴ベクトルを抽出する。このようにして抽出した判定用特徴ベクトルは、事前にコンピュータのメモリに記憶しておく。なお図１には、特にプロファイル学習用時系列データについては記載していないが、テスト用時系列データと同じルートで共起行列に変換し、その特徴ベクトルを求める。

【００３７】

次に、テスト用共起行列変換ステップでは、テストの対象となるテスト時系列データに対してウィンドウ・データ取出ステップ、スコープ・データ抽出ステップ及び共起行列変換ステップと同様のステップを実施して、テスト時系列データをテスト用共起行列に変換する。また、テスト用特徴ベクトル抽出ステップは、テスト用共起行列と固有共起行列とに基づいてテスト用時系列データについてのテスト用特徴ベクトルを抽出する。なお、テスト用特徴ベクトルを抽出する際には、図１に示すようにテスト用共起行列から平均共起行列を引いたものをベクトル化したものと先に求めた固有共起行列群をベクトル化したものとの内積を求める。

【００３８】

そして判定ステップでは、先に求めて記憶してある判定用特徴ベクトルとテスト用特徴ベクトルとに基づいて、テスト時系列データが１以上のカテゴリを含むか否かを判定する。なお判定ステップでは、具体的には、所定のベクトル識別関数を用いてテスト用時系列データと判定用特徴ベクトルとのユークリッド距離が閾値以内であるか否かによりテスト時系列データが１以上のカテゴリを含むか否か（ユーザが作成した時系列データであるか否か、すなわちユーザ以外のなりすまし者が作成した時系列データであるか否か）を判定する。

【００３９】

精度の良い異常検知システム（時系列データ異常判別方法）を構築するためには、ユーザのプロファイル（ユーザの判別用特徴ベクトル）を、コンセプチュアル・ドゥリフト（Conceptual Drift）に対応させて更新する必要がある。図４に示すような従来の方法においては、ユーザのプロファイル（ユーザの判別用特徴ベクトル）を更新する際は、識別関数における結果を利用して行う必要がある（フィードバック更新）。そのため、識別関数の結果が間違っていた場合、プロファイルが正しく更新されないという問題がある。これに対して、そこで本実施の形態では、図５に示すように、学習用の複数の時系列データ（ドメイン）に、テスト時系列データを含めて、固有共起行列群を更新する。このようにす

ると、識別関数の結果を利用せずプロファイルの更新が可能である（フィードフォーワード更新）。したがって更新を確実に行うことができる。

【００４０】

また本発明の時系列データ判定方法を用いて、コンピュータシステムに入力される時系列データの異常を判別すると、従来よりも高い精度で異常な時系列データを判別することができる。

【００４１】

Schonlau等（M.Schonlau,W.DuMouchel,W.-H.Ju,A.F.Karr,M.Theus及びY.Vardi著の「Computer intrusion Detecting masquerades」InStatｌｓticalScience,pp.16(1):58-74,2001）が提供するUNIX（登録商標）コマンドのデータを用いてなりすまし検知の実験を本実施の形態に関して行った。実験の目的は、学習用の時系列データ（ドメインデータ）のサイズの違いによる、なりすましの検知精度の違いを考察することにある。図６及び図７には、全ユーザの最初の５０個のウィンドウをドメインデータとして実験した場合を実験１として示し、同様に、全ユーザの最初の７５個のウィンドウを学習用の時系列データ（ドメインデータ）として実験した場合を実験２として示した。この実験結果からは、ドメインデータのサイズが大きい実験２の場合が、実験１よりも検知率が良いことが判った。

【００４２】

上記実施の形態では、統計的特徴抽出方法として主成分分析を用いたが、本発明の方法では主成分分析以外の他の統計的特徴抽出方法を利用できるのは勿論である。また本実施例では、識別方法として特徴ベクトルのユークリッド距離を用いたが、ユークリッド距離以外の様々なベクトル識別方法を利用できるもは勿論である。

【図面の簡単な説明】

【００４３】

【図１】複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを主成分分析法を用いて判定する本発明の時系列判定方法の実施の形態の一例を実施するためのプログラムの構成を示す図である。

【図２】複数のユーザ１乃至ユーザ３からそれぞれ送られた３つの時系列データの構成の一例を示している。

【図３】共起行列の変換を説明するために用いる図である。

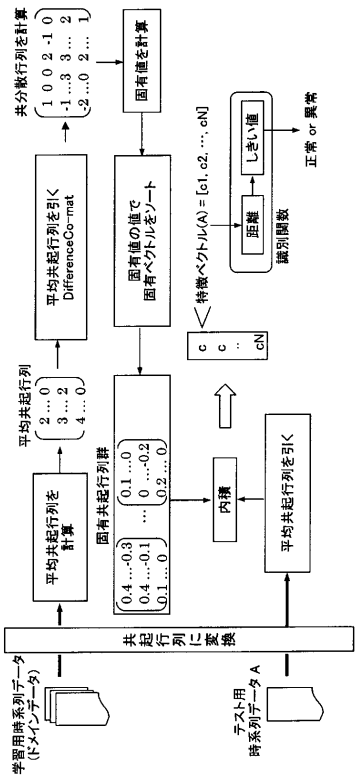
【図４】従来のプロファイル更新を説明するために用いる図である。

【図５】本発明の実施の形態によるプロファイル更新を説明するために用いる図である。

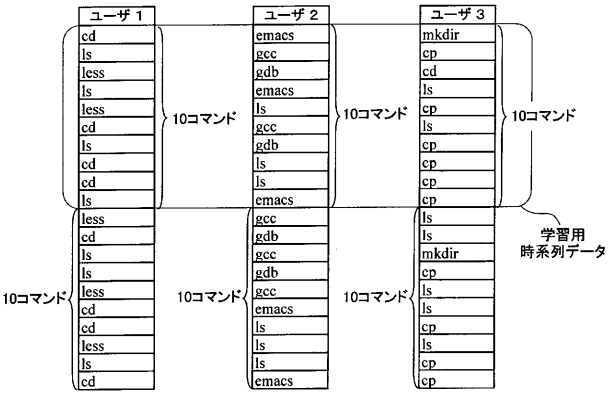
【図６】実験１及び２の基本的な概念を示す図である。

【図７】実験１及び２における検出率と誤検出の関係を示す図である。

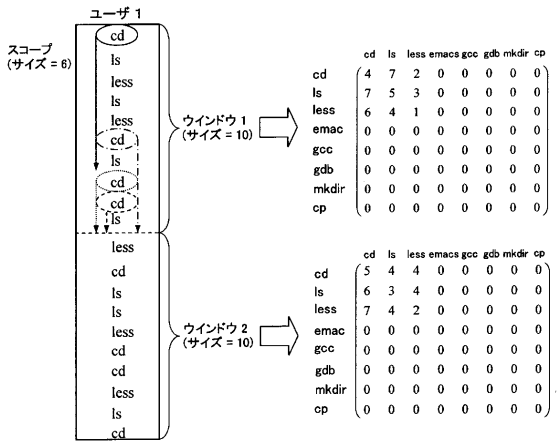
【図１】



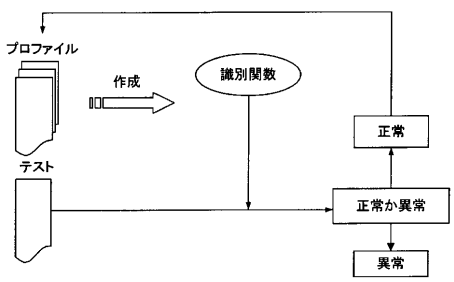
【図２】



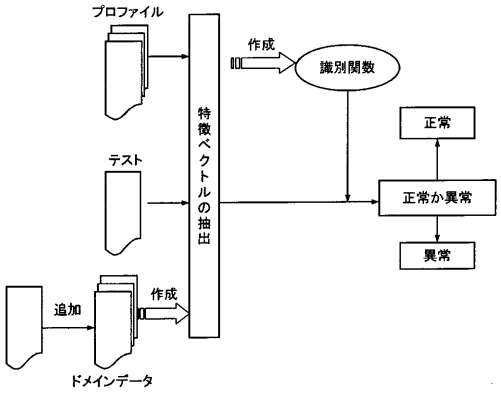
【図３】



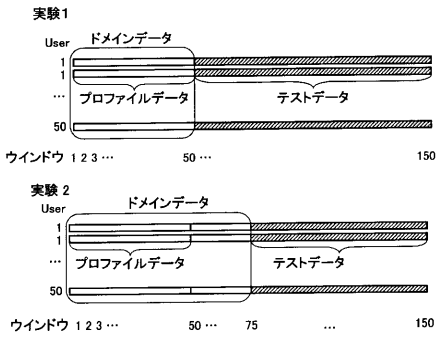
【図４】



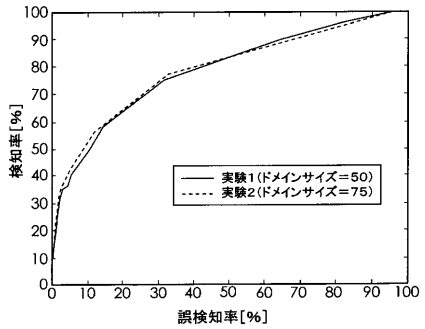
【図５】



【図６】



【図７】



【手続補正書】
【提出日】平成16年9月30日(2004.9.30)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを特徴抽出方法と識別方法とを用いて判定する時系列データ判定方法であって、
前記特徴抽出方法として、複数の時系列入力データを前記複数種類のイベントに含まれる二種類のイベント間の関連性を共起行列で表した行列データに変換したものを用いる統計的特徴抽出方法を用い、
前記識別方法として前記統計的特徴抽出方法で抽出した特徴ベクトルを識別に利用するものを用いることを実施することを特徴とする時系列データ判定方法。
【請求項２】
前記統計的特徴抽出方法が主成分分析法である請求項１に記載の時系列データ判定方法。
【請求項３】
前記複数の時系列入力データを共起行列で表した行列データに変換する際には、
前記時系列入力データをそれぞれ予め定めたデータ長のウィンドウで切り出して複数のウィンドウ・データを取り出すウィンドウ・データ取出ステップと、
前記ウィンドウ・データから前記データ長よりも短いデータ長を有する複数のスコープ

・データを時間的なずれを持って順次抽出するスコープ・データ抽出ステップと、
前記複数のウィンドウ・データを複数の前記スコープ・データに基づいて前記ウィンドウ・データに含まれる前記複数種類のイベント相互間の時系列で見た関連性の強さを示す複数の共起行列に変換する共起行列変換ステップとを実施することを特徴とする請求項１または２に記載の時系列データ判定方法。
【請求項４】
前記スコープ・データ抽出ステップでは、前記複数種類のイベントから選択した１つの種類の前記イベントが前記ウィンドウ・データに含まれる位置を基準位置として前記１つの種類のイベントに対する１以上の前記スコープ・データを抽出し、
前記共起行列変換ステップでは、前記１つの種類のイベントについての前記１以上のスコープ・データに含まれる他の１つの種類の前記イベントの数の合計値を、前記１つの種類のイベントに対する前記他の一つの種類のイベントの頻度とし、前記頻度を前記１つの種類のイベントに対する前記複数種類のイベントとの前記関連性の強さを表示する値とする変換を行って前記ウィンドウ・データを前記共起行列に変換することを特徴とする請求項３に記載の時系列データ判定方法。
【請求項５】
複数種類のイベントを含んで構成される時系列データが所定の１以上のカテゴリに属するものであるか否かを判定する時系列データ判定方法であって、
予め学習用の複数の時系列データをそれぞれ予め定めたデータ長さのウィンドウで切り出して複数のウィンドウ・データを取り出すウィンドウ・データ取出ステップと、
前記ウィンドウ・データから前記データ長よりも短いデータ長を有する複数のスコープ・データを時間的なずれを持って順次抽出するスコープ・データ抽出ステップと、
前記複数のウィンドウ・データを複数の前記スコープ・データに基づいて前記ウィンドウ・データに含まれる前記複数種類のイベント相互間の時系列で見た関連性の強さを示す複数の共起行列に変換する共起行列変換ステップと、
前記複数の共起行列を入力として統計的特徴抽出方法により特徴ベクトルを求めるための基礎となる固有共起行列群を決定する固有共起行列群決定ステップと、
前記１以上のカテゴリを含む１以上のプロファイル学習用時系列データに対して前記ウィンドウ・データ取出ステップ、前記スコープ・データ抽出ステップ及び前記共起行列変換ステップと同様のステップをそれぞれ実施して、前記１以上のプロファイル学習用時系列データを１以上のプロファイル用共起行列に変換するプロファイル用共起行列変換ステップと、
前記１以上のプロファイル用共起行列と前記固有共起行列群とに基づいて前記１以上のプロファイル学習用時系列データについての１以上の判定用特徴ベクトルを抽出する判定用特徴ベクトル抽出ステップと、
テストの対象となるテスト時系列データに対して前記ウィンドウ・データ取出ステップ、前記スコープ・データ抽出ステップ及び前記共起行列変換ステップと同様のステップを実施して、前記テスト時系列データをテスト用共起行列に変換するテスト用共起行列変換ステップと、
前記テスト用共起行列と前記固有共起行列群とに基づいて前記テスト用時系列データについてのテスト用特徴ベクトルを抽出するテスト用特徴ベクトル抽出ステップと、
前記１以上の判定用特徴ベクトルと前記テスト用特徴ベクトルとに基づいて、前記テスト時系列データが前記１以上のカテゴリに属するか否かを判定する判定ステップとからなる時系列データ判定方法。
【請求項６】
前記スコープ・データ抽出ステップでは、前記複数種類のイベントから選択した１つの種類の前記イベントが前記ウィンドウ・データに含まれる位置を基準位置として前記１つの種類のイベントに対する１以上の前記スコープ・データを抽出し、
前記共起行列変換ステップでは、前記１つの種類のイベントについての前記１以上のスコープ・データに含まれる他の１つの種類の前記イベントの数の合計値を、前記１つの種

類のイベントに対する前記他の一つの種類のイベントの頻度とし、前記頻度を前記１つの種類のイベントに対する前記他の種類のイベントの前記関連性の強さを表示する値とする変換を行って前記ウィンドウ・データを前記共起行列に変換することを特徴とする請求項５に記載の時系列データ判定方法。

【請求項７】

前記判定用特徴ベクトル抽出ステップでは、前記プロファイル用共起行列と前記固有共起行列群とをベクトル化した後にその内積を求めて前記判定用特徴ベクトルを決定し、

前記テスト用特徴ベクトル抽出ステップでは、前記テスト用共起行列と前記固有共起行列群とをベクトル化した後にその内積を求めて前記テスト用特徴ベクトルを抽出することを特徴とする請求項５に記載の時系列データ判定方法。

【請求項８】

前記判定ステップでは、所定のベクトル識別関数を用いて前記テスト用時系列データと前記判定用特徴ベクトルとのユークリッド距離が閾値以内であるか否かにより前記テスト時系列データが前記１以上のカテゴリに属するか否かを判定する請求項５に記載の時系列データ判定方法。

【請求項９】

前記学習用の複数の時系列データに、前記テスト時系列データを含めて、前記固有共起行列群を更新することを特徴とする請求項５に記載の時系列データ判定方法

【請求項１０】

請求項１乃至９のいずれか１項に記載の時系列データ判定方法を用いて、コンピュータシステムに入力される時系列データの異常を判別することを特徴とする時系列データ異常判別方法。