2023/03/15

統計的パターン認識とともに
基礎と医学問題への応用

山口大学大学院創成科学研究科
浜本　義彦
hamamoto@yamaguchi-u.ac.jp

内容

第1部　統計的パターン認識の基礎

(1) ﾊﾟﾀｰﾝ認識とは

(2) ﾊﾟﾀｰﾝ認識の困難さ

①ｸﾗｽの存在性
（学習の仕方、ｻﾝﾌﾟﾙとは、帰納的学習の宿命である不確実性への対処方法）
②観測のあり方（ｻﾝﾌﾟﾙからの観測ﾃﾞｰﾀの獲得方法）
③役に立つ観測ﾃﾞｰﾀとは（特徴選択の重要性）
④価値観と認識（ﾊﾟﾀｰﾝ認識の本質）

第2部 医学問題への応用

(1)　医学問題

(2) 離散Bayes識別則

(3) 個別化医療への適用

第1部　統計的パターン認識の基礎

ﾊﾟﾀｰﾝ認識とは

「見る」と「見なす」の違い

ﾊﾟﾀｰﾝ認識では、「何を何と見なすか」という対応付け問題に回答する。
この回答が認識結果である。

例えば、ｺﾝﾋﾟｭｰﾀは目の前に存在する本を見て、これは無数に存在する本の一つの事例であって、本というｸﾗｽの一員である、と見なす。

本というｸﾗｽは、ﾋﾄの間で共通の概念としてどのように形成されているのか？
（ｸﾗｽの定義は如何に?）

ｺﾝﾋﾟｭｰﾀによるﾊﾟﾀｰﾝ認識とは
そもそもｺﾝﾋﾟｭｰﾀに本を見せるとはどういうことなのか？　仮に、本に関するﾃﾞｰﾀを獲得し、それをｺﾝﾋﾟｭｰﾀに入力すること、とする。そうであれば、本から如何なるﾃﾞｰﾀを獲得すれば、ｺﾝﾋﾟｭｰﾀは本という概念を形成できるのか？

命題
何も考えずに勝手に獲得したﾃﾞｰﾀを用いて、本の概念がｺﾝﾋﾟｭｰﾀの中に形成されるのか

ﾊﾟﾀｰﾝ認識の定義

外界に存在する認識対象を、それが属すべきｸﾗｽへ対応づける機能
認識対象に関する条件：観測可能であり、自己同一性がある

ｸﾗｽとは？

認識主体者（認識問題の定義者）が、同じとみなす認識対象の集団に対する概念（例えば、病名、本、机、犬等、ﾋﾄが創作した概念）

ｺﾝﾋﾟｭｰﾀによるﾊﾟﾀｰﾝ認識

認識対象を観測して複数のﾃﾞｰﾀを獲得する。次に、ﾃﾞｰﾀの組を用いて認識対象をｺﾝﾋﾟｭｰﾀ内のﾊﾟﾀｰﾝ（ﾊﾟﾀｰﾝﾍﾞｸﾄﾙ）として表現する。
つまり、ｺﾝﾋﾟｭｰﾀ外の認識対象とｺﾝﾋﾟｭｰﾀ内のﾊﾟﾀｰﾝが同一視される。
ﾊﾟﾀｰﾝとは、あるｸﾗｽの一員としての個体である。

このﾊﾟﾀｰﾝを、それが属すべきｸﾗｽに対応づけることで、そのｸﾗｽの一員とみなす。
これが、**ｺﾝﾋﾟｭｰﾀによるﾊﾟﾀｰﾝ認識**である。

ﾊﾟﾀｰﾝ認識の実際

文字認識

数字認識では、人間が書いた、または機械で印刷した幾何学図形（認識対象）をｶﾒﾗ（観測器）で撮って、その幾何学図形が１０個の数字（ｸﾗｽ）のどれに該当するのかをｺﾝﾋﾟｭｰﾀが読み取る。

顔認識

ｶﾒﾗ（観測器）に映った人間の顔（認識対象）が、予め登録された顔（ｸﾗｽ）の中の誰であるかをｺﾝﾋﾟｭｰﾀが言い当てる。

病気の診断

・患者（認識対象）を検査（観測）し、その検査ﾃﾞｰﾀを見て、患者がどの病気（ｸﾗｽ）に罹っているのかを、ｺﾝﾋﾟｭｰﾀが診断する。
・画像診断では、患者（認識対象）をCT装置（観測器）で撮影した放射線画像をｺﾝﾋﾟｭｰﾀが読影して、病変の有無（ｸﾗｽ）を診断する。

音声認識

ﾋﾄの声（認識対象）をﾏｲｸﾛﾎﾝ（観測器）で捉え、その声が何である（ｸﾗｽ）かを、ｺﾝﾋﾟｭｰﾀが聞き取る。

ﾊﾟﾀｰﾝ認識の困難さ

①ｸﾗｽは存在するのか？
もし存在するとすれば、ｸﾗｽをどのようにして　ｺﾝﾋﾟｭｰﾀに教えることができるのか？　（学習の仕方）

②観測系の在り方は如何にあるべきか？
認識対象（ｻﾝﾌﾟﾙ）から、如何なるﾃﾞｰﾀを獲得すべきか？

③ﾃﾞｰﾀは選ばれるべきものなのか？
獲得されたﾃﾞｰﾀをそのまま使うのか、それとも取捨選択するのか？

①ｸﾗｽは存在するのか

教師とは？

何を何と認識する（みなす）のかという認識問題を定義する認識主体者であって、事例となるｻﾝﾌﾟﾙに正解のｸﾗｽ名を付与する者

例　病気の診断では「医師」が教師であり、ｻﾝﾌﾟﾙ（患者）にｸﾗｽ（病名）を与える（診断）

教師あり学習

統計学では、外的基準（教師）のある統計的推測であり、多変量解析の判別分析等が該当

Pattern Recognition　ﾊﾟﾀｰﾝの再認識

認識主体者である教師が、ｸﾗｽを定義する。ｺﾝﾋﾟｭｰﾀは、外界の認識対象と同一視されるﾊﾟﾀｰﾝがどのクラスの一員であるかを一度学習しておく。そして、新たなﾊﾟﾀｰﾝがどのｸﾗｽの一員であるかを問われ、再び認識する。

つまり、ﾊﾟﾀｰﾝ認識では、ﾊﾟﾀｰﾝをｸﾗｽに対応づけるために、予め（学習前に）ｸﾗｽが定まっていることが前提となる。つまりｸﾗｽが存在しないと、ﾊﾟﾀｰﾝ認識はできない。

ｺﾝﾋﾟｭｰﾀにｸﾗｽ（概念）をどのようにして教えるのか？（概念の形成は如何に？）

文字認識では、ｺﾝﾋﾟｭｰﾀが認識対象である数字「２」を表す幾何学図形を観測して、それを数字「２」のｸﾗｽの一員としてみなす。では数字「２」という概念をﾋﾄはどのように形成してﾋﾄ同士で共通理解しているのであろうか？

ｺﾝﾋﾟｭｰﾀに正解のｸﾗｽ名付き事例（ｻﾝﾌﾟﾙ）を見せて、事例間に存在する規則性を「帰納的」に学習させる。

この学習を教師あり学習と呼び、帰納的推論の一種に該当する。

しかし、ｻﾝﾌﾟﾙを用いた学習は、例外なく「帰納的推論の宿命」から逃げられない。

論点Ⅰ：不確実性

ｻﾝﾌﾟﾙを用いた学習とは？

||

帰納的推論

宿命

①絶対に正しいことが保証されない。

②ｻﾝﾌﾟﾙによって学習結果が変わる。

帰納的推論を基に識別を行うと、誤りの可能性がある。

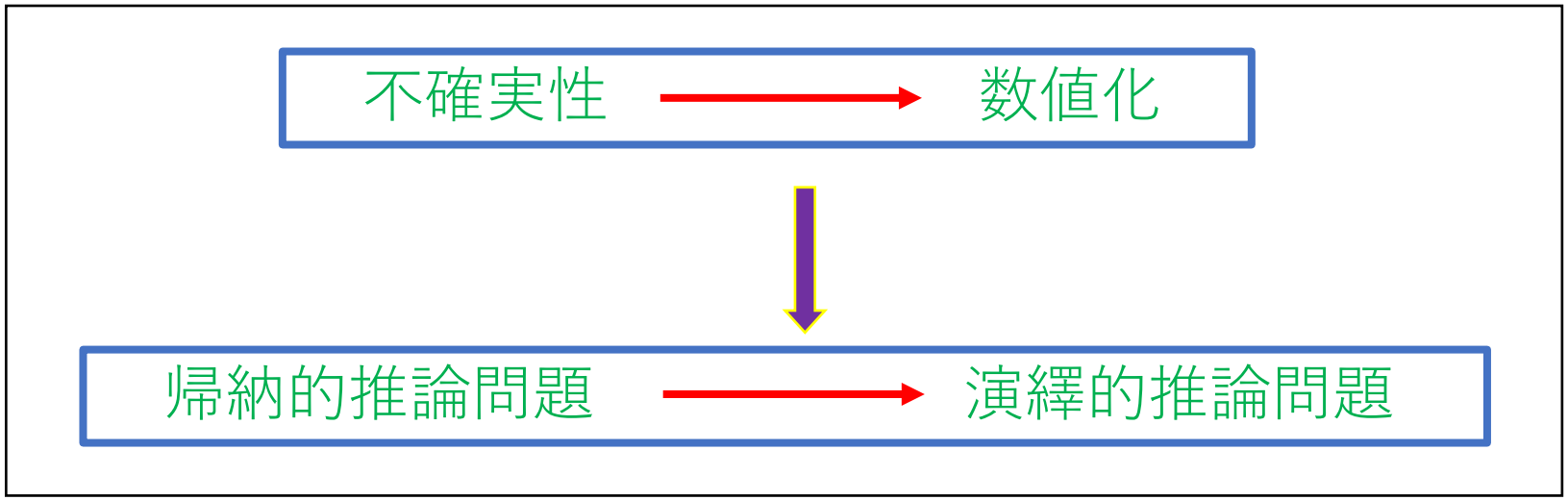
もし誤りが原理的に避けられないのであれば、いかなる識別則を採用すべきか？

誤りに関与する不確実性をｺﾝﾄﾛｰﾙして、誤りを最小にする識別則を用いるべき

統計学とは、不確実性を科学する学問である（C. R. Rao）

不確実性は、原因と結果との間に一対一対応が欠如していることで生じる。一つの結果に対して複数の原因が考えられる。結果を見て複数の原因の中から真の原因を言い当てる推論、帰納的推論を行う。帰納的推論には絶対はない。帰納的推論は、演繹的推論と異なり、推論に正確さが欠ける。これが不確実性である。この不確実性に対処するため、統計学は不確実性の数値化を導入し、解くべき問題を変換した。

統計学の考え



統計的ﾊﾟﾀｰﾝ認識とは

不確実性に対処するために統計学のｱﾌﾟﾛｰﾁを採用する

ﾊﾟﾀｰﾝ認識には、生体に学ぶ立場（ﾆｭｰﾗﾙﾈｯﾄ、ﾃﾞｨｰﾌﾟﾗｰﾆﾝｸﾞ）や純粋に数学の幾何学問題として解く立場（ｻﾎﾟｰﾄﾍﾞｸﾀｰﾏｼﾝ）等の非統計的手法がある。

統計的ﾊﾟﾀｰﾝ認識は、1950年代に多変量解析を母体として誕生して独自に発展した、不確実性に対処できるﾊﾟﾀｰﾝ認識の一つである。

統計的ﾊﾟﾀｰﾝ認識のBayes識別則は、
①不確実性を事後確率の形で数値化し、
②事後確率を用いて識別問題を、帰納的推論から演繹的推論の枠組みへ変換して解く。つまり不確実性のある帰納的推論問題を、不確実性を表す数値の大小の比較という演繹的推論（数学）問題に変換して解く。

Bayes識別則により、理論上、誤識別率は最小化される。しかし、現実は事後確率は未知で、その推定値（誤差あり）を使うことになり、最小化は実現されない。

ﾊﾟﾀｰﾝ認識の困難さ

①ｸﾗｽは存在するのか？

もし存在するとすれば、ｸﾗｽをどのようにして　ｺﾝﾋﾟｭｰﾀに教えることができるのか？　（学習の仕方）

②観測系の在り方は如何にあるべきか？

認識対象（ｻﾝﾌﾟﾙ）から、如何なるﾃﾞｰﾀを獲得すべきか？

③ﾃﾞｰﾀは選ばれるべきものなのか？

獲得されたﾃﾞｰﾀをそのまま使うのか、それとも取捨選択するのか？

②観測系の在り方

認識対象から、如何なるﾃﾞｰﾀを獲得すべきか？
（ｺﾝﾋﾟｭｰﾀに認識対象を見せるとは？）

認識対象に関する知識が完全である（非現実　不確実性のない場合）

例　２次方程式の識別問題
　認識対象：２次方程式
　ｸﾗｽ　：　実根をもつ２次方程式と虚根をもつ２次方程式
　観測：　認識対象から如何なるﾃﾞｰﾀを獲得すべきか？

　答え：　判別式（特徴）の値

これにより１００％の精度で２次方程式を識別でき、そこに不確実性はない
（非負なら必ず実根、負なら必ず虚根）。

これが可能なのは、認識対象である２次方程式の理論が解明されているからである。

認識対象に関する知識は不完全である（現実）

例えば、文字認識では技術者が試行錯誤で創意工夫したﾃﾞｰﾀを獲得しているが、それが良いという保証はなく、決め手となるﾃﾞｰﾀは見つかっていない。

では、どうすれば良いのか？

原則

認識対象の研究に基づいてﾃﾞｰﾀを獲得する。

例：患者が認識対象であれば、患者に関する研究とは「医学」である。
医学によって、患者に対して何を検査すべきかが決まる。

認識対象によって、観測手段が異なる。例えば、顔や文字の認識であれば、ｶﾒﾗによって認識対象を観測し、画像ﾃﾞｰﾀを得る。病気診断では医師が検査項目を適宜決める。何を観測するかは認識対象の研究を踏まえるべきで、認識対象の影響を受ける。

従って「ﾊﾟﾀｰﾝ認識は、必然的に、認識対象に依存した個別論（文字認識、音声認識等）」となる。

注意すべきは、どんなﾃﾞｰﾀを獲得すべきか、やってみないと分からない、試行錯誤が不可欠。

ﾃﾞｰﾀの獲得と解析

ﾃﾞｰﾀ獲得では、まず研究ﾃﾞｻﾞｲﾝの段階で医師が仮説を立て、次に医師とﾃﾞｰﾀｻｲｴﾝﾃｨｽﾄが一緒になって仮説検証のためのﾃﾞｰﾀ獲得について十分な協議を行う。これを踏まえて、ﾃﾞｰﾀｻｲｴﾝﾃｨｽﾄが解析法を検討する。しかし、多くは一切の相談なしにﾃﾞｰﾀを獲得してしまうので、ﾃﾞｰﾀｻｲｴﾝﾃｨｽﾄから価値のある結果は期待できない。

実際は認識対象から有望と思われる
ﾃﾞｰﾀを可能な限り獲得する

→

情報爆発
ﾋﾞｯｸﾞﾃﾞｰﾀ誕生

例：がんは遺伝子異常の病気である。
専門家の思い：たぶん遺伝子を調べれば、がんは解明されるはず（確たる根拠はない）

ﾋﾞｯｸﾞﾃﾞｰﾀの例

分子生命科学の発展により遺伝子変異など、ﾍﾟﾀｽｹｰﾙの膨大なﾃﾞｰﾀが簡単に獲得可能となり、20年前に**網羅的解析が大流行した。**

膨大なﾃﾞｰﾀは、人間の手に負えない
そこでｺﾝﾋﾟｭｰﾀ，AIの登場

ﾊﾟﾀｰﾝ認識の困難さ

①ｸﾗｽは存在するのか？

もし存在するとすれば、ｸﾗｽをどのようにして　ｺﾝﾋﾟｭｰﾀに教えることができるのか？　（学習の仕方）

②観測系の在り方は如何にあるべきか？

認識対象（ｻﾝﾌﾟﾙ）から、如何なるﾃﾞｰﾀを獲得すべきか？

③ﾃﾞｰﾀは選ばれるべきものなのか？

獲得されたﾃﾞｰﾀをそのまま使うのか、それとも取捨選択するのか？

③ﾃﾞｰﾀは選ばれるべきものなのか？

統計的ﾊﾟﾀｰﾝ認識の研究成果から

ﾃﾞｰﾀには
識別に役に立つﾃﾞｰﾀと役に立たないﾃﾞｰﾀがある。つまり、
ﾃﾞｰﾀには重要性の度合いがある。

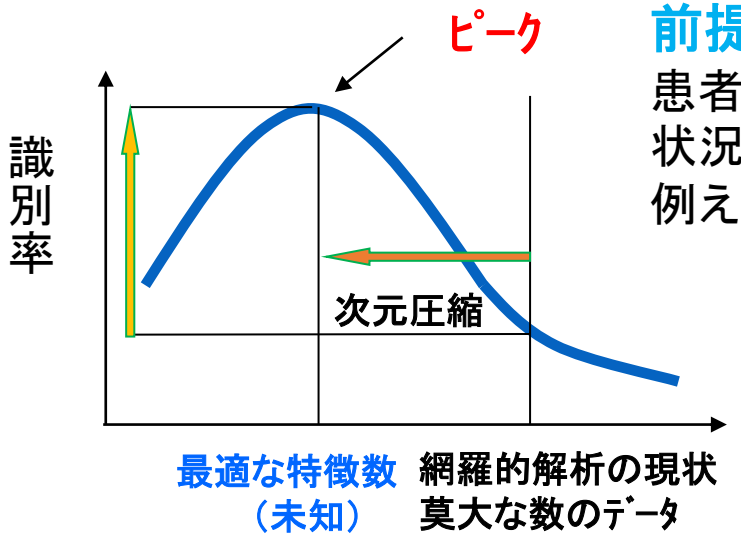
たとえ役に立たなくても、
それが悪さをしなければ、無視すれば良い。

しかし、
役に立たないﾃﾞｰﾀが、識別において悪さをするのである。

従って、役に立たないﾃﾞｰﾀを削除し、
役に立つﾃﾞｰﾀだけを用いなければならない。

特徴選択の必要性

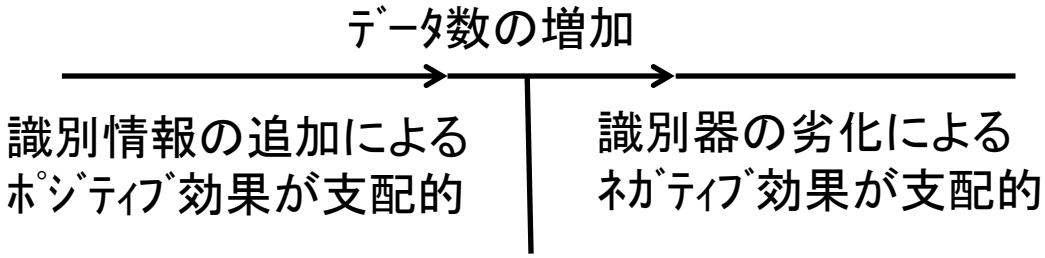
ﾋﾟｰｷﾝｸﾞ現象



前提：ｻﾝﾌﾟﾙ数一定

患者数を増やすことができず、ﾃﾞｰﾀ数だけを増やす状況。

例えば遺伝子発現やﾒﾁﾙ化のﾃﾞｰﾀを用いる場合等。



ﾃﾞｰﾀ（ﾏｰｶｰ）数を増やすと、情報が追加されるため最初は識別性能は向上するが、やがて飽和し、更に増加すると識別器の学習が不十分になり、**識別性能が劣化し始める**。そのため、識別には全てのﾃﾞｰﾀではなく、重要なﾃﾞｰﾀだけを選んで使うべきである。

特徴（標的ﾏｰｶｰ）の組合せは重要である

１．Elashoffらの成果（１９６７年）

Elashoff et al., Biometrika, 54, pp.668-670

特徴は独立

	誤識別率（%）
特徴F_1	7
特徴F_2	14
特徴F_3	21
F_1，F_2	6
F_1，F_3	4

単独で評価して良いものを組合せても
良いとは限らない。

２．Coverらの成果（１９７７年）

Cover et al., IEEE Trans. SMC, 7, 9, pp.657-661

全ての組合せを調べないと最適解を見つける
ことはできない。しかし計算不可能となる。
実際は、計算可能な部分的探索を行う。

ﾋﾄは一つか、同時には二つのﾏｰｶｰの組合せ
しか理解できない。三つ以上のﾏｰｶｰは無理。

ﾎﾟｲﾝﾄ
特徴（標的ﾏｰｶｰ）の組合せが重要

１次元思考（人手）から
多次元思考（ｺﾝﾋﾟｭｰﾀ）へ

論点Ⅱ：価値観と認識

ﾃﾞｰﾀの重要性は、ﾃﾞｰﾀの質によって問われる。

質は、認識主体者の判断、関心の有無などの価値観に基づき、決して論理的なものではない。

①医師の診断では、想定される疾患にとって重要な検査とそうではない検査がある。

②分子生命科学者は、ある種のがんに関与する遺伝子群があり、全ての遺伝子が等しく平等に関与するわけではないと考えている。

③象とｸｼﾞﾗとﾏｸﾞﾛに対し、海洋学者には海に住むというﾃﾞｰﾀからｸｼﾞﾗとﾏｸﾞﾛは似ていて、動物学者にとっては哺乳類というﾃﾞｰﾀから象とｸｼﾞﾗが似ている。

質は価値観に基づき、価値観はﾋﾄが定める。ﾋﾄは、ﾋﾄが意味がある、価値があると認めたｸﾗｽを定義（創作）する。

価値観は、論理的ではなく、非論理的である。

従って、ﾊﾟﾀｰﾝ認識には非論理的要素が必要である。

①学習におけるｻﾝﾌﾟﾙの選び方

当該分野の専門家（医師）が、ｸﾗｽを代表し得る典型的なｻﾝﾌﾟﾙを選ぶ。
ﾊﾟﾀｰﾝ認識では「学習に相応しい、質の高いｻﾝﾌﾟﾙ」を、専門家が「無作為」ではなく意図をもって選ぶ。必ずしも数が多ければよい、というわけではない。

②ｻﾝﾌﾟﾙからのﾃﾞｰﾀ獲得の仕方

何も考えずに勝手にﾃﾞｰﾀを獲得するのではなく
「ﾋﾄが定めたｸﾗｽを形成し得るように、よく考えて選ばれたﾃﾞｰﾀを獲得する。」

ﾃﾞｰﾀはｸﾗｽと不可分の関係にある（決して両者を分離して議論すべきではない）。

ﾊﾟﾀｰﾝ認識において、非論理的要素であるﾋﾄの価値観の導入が不可欠であることを初めて説いたのは、渡辺慧である。
（みにくいｱﾋﾙの子定理　60年前）

統計学は、数学の一分野に属し、論理的である。原理的に、論理（客観）と非論理（主観）とは相容れない。統計学が科学たるためには、主観を脱し、客観化が絶対条件となる。このため、統計学は、質の議論を意識的に避け、数のみを論じる。しかも無作為化を有効とするために極めて多数のｻﾝﾌﾟﾙを要求する。

だが、現実はｻﾝﾌﾟﾙ数が少なく、統計学の要求と矛盾し、理論は破綻する。

ここにSmall Sample Size問題が根源的な問題となり、その深刻さを次元数の増加が格段に高める。

統計学と統計的ﾊﾟﾀｰﾝ認識とは
似て　　不確実性の数値化
非なる　　価値観の導入
関係にある。

第2部 医学問題への応用

(1) 医学問題

(2) 離散Bayes識別則

(3) 個別化医療への適用

用語の定義

ｸﾗｽ(医学的概念, 例えば病名), ｻﾝﾌﾟﾙ(患者), 観測(検査), 観測項目(検査項目=ﾏｰｶｰ),

観測ﾃﾞｰﾀ(検査ﾃﾞｰﾀ=ﾏｰｶｰの測定値), 特徴(有用な観測項目=標的ﾏｰｶｰ), 教師(医師)

AIによる診療の支援とは

立場：　あくまで主役は医師、ｺﾝﾋﾟｭｰﾀ（AI)は支援のみ

診療におけるAIの活用とは、ｺﾝﾋﾟｭｰﾀが医師の気づかない点を指摘しﾋﾝﾄを与え、医師の臨床推論力や発想力を高めることである。

盲点

ｻﾝﾌﾟﾙ（症例）の収集やｻﾝﾌﾟﾙからのﾃﾞｰﾀ獲得については、機械学習は何も言えない。
同様に、多変量解析も何も答えない。単に与えられたﾃﾞｰﾀを処理するだけである。

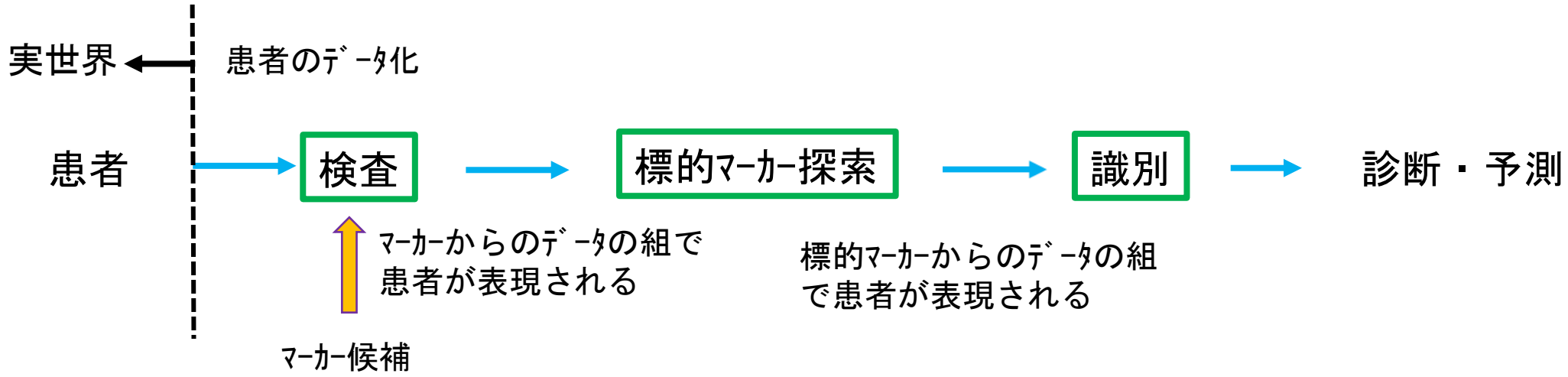
そもそもﾃﾞｰﾀに必要な情報が含まれているか否かは、
解析前は不明（やってみないと分からない）。しかも何ら理論的道すべもない。

実際はｻﾝﾌﾟﾙの収集、ﾃﾞｰﾀの獲得が一番重要である。にもかかわらず、軽視されている。
どの教科書にも書かれていない。結局、経験からでしか学べない。

医学問題では、医師のみが医学知識をもって議論できる。

医師は、ｺﾝﾋﾟｭｰﾀ（AI)の結果を決して鵜呑みにせず、医学的に解釈すること

医学問題におけるﾊﾟﾀｰﾝ認識ﾓﾃﾞﾙ



標的ﾏｰｶｰ探索と患者層別化

患者に対して検査を行い、検査項目(ﾏｰｶｰ)毎に検査ﾃﾞｰﾀを獲得し、検査ﾃﾞｰﾀの組を用いて患者をﾊﾟﾀｰﾝと記述する。ﾊﾟﾀｰﾝを、予め定めたｸﾗｽ、例えば薬剤効果の有無、予後の良好不良等に患者毎に識別する。これを患者層別化という。識別に役立つﾏｰｶｰを標的ﾏｰｶｰと呼び、識別精度は標的ﾏｰｶｰに依存する。

ﾊﾟﾀｰﾝ認識の困難さ（整理）

①学習は、帰納的推論によるため、識別の誤りを原理的にｾﾞﾛにすることはできない。
②ﾏｰｶｰ候補の選定は重要である。病態が未解明の場合、患者から予め何をﾃﾞｰﾀとして取れば良いかは不明であるため、網羅的にﾏｰｶｰ候補を用意する。するとﾃﾞｰﾀ数（次元数）は増大して高次元となる。つまり「情報の爆発」である。このため、標的ﾏｰｶｰ探索（特徴選択）が必須となる。
③医学分野では質の高いｻﾝﾌﾟﾙ（症例）は少ない。ｻﾝﾌﾟﾙの質は、医師によって医学的に評価される。

統計的ﾊﾟﾀｰﾝ認識と標的ﾏｰｶｰ探索、患者層別化との関係

統計的ﾊﾟﾀｰﾝ認識の定義
①観測：認識対象（患者）から観測（ﾏｰｶｰ）によってﾃﾞｰﾀ（測定値）の獲得
②特徴選択：観測（ﾏｰｶｰ）の中から認識に有用な特徴（標的ﾏｰｶｰ）の組合せの選択
③識別：特徴（標的ﾏｰｶｰ）の組を用いて認識対象（患者）の、それが属すべきｸﾗｽ
（ﾚｽﾎﾟﾝﾀﾞｰ、ﾉﾝﾚｽﾎﾟﾝﾀﾞｰのいずれか）への識別

⇩

解くべき問題の変換

医学問題		工学問題
(1) 標的ﾏｰｶｰ探索問題	⇒	特徴選択問題
(2) 患者層別化問題	⇒	認識対象の識別問題

この二つの問題を解くには、実績のある統計的ﾊﾟﾀｰﾝ認識に基づいて

1) **特徴選択技術**により標的ﾏｰｶｰの組合せを探索し、
2) 標的ﾏｰｶｰの組を用いた**識別技術**により患者をﾚｽﾎﾟﾝﾀﾞｰかﾉﾝﾚｽﾎﾟﾝﾀﾞｰかに識別する

ことになる。しかし事はそう簡単ではなく、越えるべき大きな壁がある。

越えるべき大きな壁と解決策

ｾﾝｼﾝｸﾞ技術の発展により次世代ｼｰｹﾝｻｰからの遺伝子変異等、各種ｵﾐｯｸｽ情報のﾃﾞｰﾀを獲得する技術は確立された。しかし、患者層別化（標的ﾏｰｶｰ探索を含む）を目的とした、ﾀｲﾌﾟの異なる膨大なﾃﾞｰﾀを同じ枠組みで一括して処理できる解析法はない。

ﾌﾞﾚｲｸｽﾙｰその1

①ﾃﾞｰﾀの「こと」(記号)化

②「こと」を処理できる解析法 ⇦ 独自の離散Bayes識別則と離散系最近傍解析

数値ﾃﾞｰﾀだけでなく、医学上重要な記号ﾃﾞｰﾀ(遺伝子変異等)も混在する臨床ﾃﾞｰﾀを一括解析できる。

ﾌﾞﾚｲｸｽﾙｰその2

Lancet論文[doi.org/10.1016/S0140-6736(03)12775-4]から進化した独自の特徴選択技術

ｲﾝﾊﾟｸﾄﾌｧｸﾀｰ18.316(2003年当時)　被引用件数443回(Scopus)

がんは特にｻﾝﾌﾟﾙ（症例）数が少なく、しかしﾃﾞｰﾀ（次元）数は莫大で、正に「情報の爆発」である。

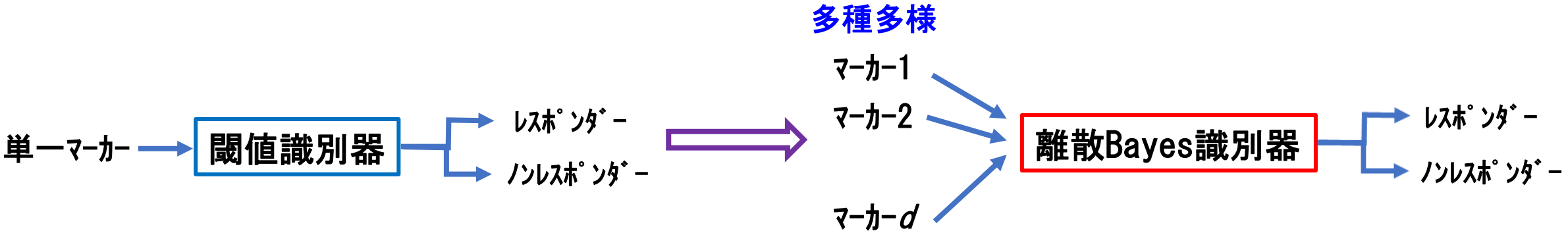
Small Sample Size and High Dimensionalityの状況下でも、ｻﾝﾌﾟﾙ変動にﾛﾊﾞｽﾄな（汎化性の高い）特徴選択（標的ﾏｰｶｰ探索）が可能。

特長

① ﾏｰｶｰの測定値であるﾃﾞｰﾀを全て記号ﾃﾞｰﾀとして「こと（医学的概念）」（言葉）化
（数値ﾃﾞｰﾀはcut-off値により離散化、記号ﾃﾞｰﾀは必要に応じて適度な粗さに再記号ﾃﾞｰﾀ化）

② 記号ﾃﾞｰﾀとしての「こと」を扱える独自の離散Bayes識別則に基づき、統計的ﾊﾟﾀｰﾝ認識の特徴選択と識別の両方が行える解析法の確立

③ 数値ﾃﾞｰﾀと記号ﾃﾞｰﾀのようにﾀｲﾌﾟが異なり、多種多様で莫大な数のﾏｰｶｰを候補とし、その中から有力な標的ﾏｰｶｰの組合せを、高速に、網羅的に探索する。
（標的ﾏｰｶｰを個々に考えるのではなく、標的ﾏｰｶｰの組合せに着目する）



従来の識別（単一ﾏｰｶｰでは不完全）

新AI技術の識別（標的ﾏｰｶｰの組合せ利用）
個々のﾏｰｶｰが不完全であっても、それらを組合せて互いに補うことで不完全さを低減させ、識別の高精度化が図れる

個人情報保護法におけるﾃﾞｰﾀの第三者提供（AI開発に必須）

本人の同意なしで目的外利用の第三者提供は可能か
ﾃﾞｰﾀ利用：提供元（病院等）では当初の目的は患者本人の診療で、提供先（企業）のAI開発は目的外利用

個人情報保護法（平成27年改正）

対策1：匿名加工情報

背景：医療情報は要配慮個人情報である。このためｵﾌﾟﾄｱｳﾄは不可。
このままでは目的外利用の第三者提供はできない。

特徴：特定の個人が識別できない情報に加工する（非個人情報化）

欠点：提供元がどこまで加工しなければならないのかは不明。もし不完全であれば提供元に責任が残る。
更に法のもとで定まった手続きが複雑である。非現実的

次世代医療基盤法（平成29年）

（医療に特化した、個人情報保護法の特別法）

対策2：匿名加工医療情報

背景：病院側に責任がなく、企業からは対策1よりも手続きがもっと簡便な加工法が求められた。

特徴：大臣認定事業者が全ての加工と責任を引き受ける
（非個人情報化）

欠点：加工について口を出せない。
そのため、AI開発においてﾃﾞｰﾀが使い物にならないﾘｽｸがある。

個人情報保護法（令和2年改正）

（令和4年4月施行）

対策3：仮名加工情報

背景：企業からは、対策1より使い易く、対策2より自由度のある加工法が求められた。

特徴：特定の個人が、単独の情報では識別できないが、他の情報と照合すれば識別できる加工。敢えて不完全さ（完全性で責任が問われないという解釈）を利点とする。

欠点：対策1，2と異なり、第三者提供はできない。但し、病院からの委託、又は病院との共同利用という病院の管理下であれば第三者提供、AI開発が可能（欠点というより制約）

個人情報の利活用促進（規制緩和）の流れ

離散Bayes識別則のｱﾌﾟﾛｰﾁは、既存の数値を対象とした解析法とはまったく異なり、対策1,2,3は不要

個別化医療における標的ﾏｰｶｰ探索と患者層別化に特化した離散Bayes識別則の研究成果1：肝がんの早期再発の予測

背景

肝がんの怖さは高い再発率。完全切除しても術後1年以内に約30%が再発する。再発を予測できれば効果的な先制医療を実施でき、もし無再発ならば副作用のある抗がん剤やCT検査が不要となり、医療費の抑制にもなる。

医学問題

手術で肝がんを完全切除した患者を対象に、容易に入手できる保険適用の臨床ﾃﾞｰﾀを用いて術後1年以内の早期再発を予測せよ。

工学問題

医師が選定した11ﾏｰｶｰ候補の中から離散Bayes識別則により標的ﾏｰｶｰを選択し、早期再発を予測せよ。

結果

選択された5標的ﾏｰｶｰの組合せは、腫瘍数、腫瘍ｻｲｽﾞ, ICG（数値ﾃﾞｰﾀ), vp, Liver damage（記号ﾃﾞｰﾀ）の混在であり、**ﾃｽﾄｻﾝﾌﾟﾙに対して感度86%（癌再発の検出率）, 特異度49%を得た。** 良く知られているTNM分類やModified JISよりも高感度で, ROC解析でも有効性を示した。

Prediction of early recurrence of liver cancer by a novel discrete Bayes decision rule for personalized medicine, *BioMed Research International*, 2016. doi.org/10.1155/2016/8567479

離散Bayes識別則の研究成果2: 早期胃がんのﾘﾝﾊﾟ節転移の診断

背景

早期胃がんは内視鏡的治療で対応できるが、ﾘﾝﾊﾟ節転移が疑われると外科手術となる。しかし実際にはﾘﾝﾊﾟ節転移のない患者にも不要な外科手術が行われている。

医学問題

早期胃がんに対する内視鏡的治療後に、容易に入手できる保険適用の臨床ﾃﾞｰﾀを用いて、ﾘﾝﾊﾟ節転移を診断せよ。

工学問題

医師が選定した8ﾏｰｶｰ候補の中から離散Bayes識別則により標的ﾏｰｶｰを選択し、転移を診断せよ。

結果

選択された3標的ﾏｰｶｰの組合せは、深達度、ﾘﾝﾊﾟ管侵襲、静脈侵襲（全て記号ﾃﾞｰﾀ）であり、
ﾃｽﾄｻﾝﾌﾟﾙに対して感度100%(転移の見逃しｾﾞﾛ)，特異度86%を得た。

Lymph node metastasis can be determined by just tumor depth and lymphovascular invasion in early gastric cancer patients after endoscopic submucosal dissection,
***European Journal of Gastroenterology and Hepatology*, 2017. doi: 10.1097/MEG.0000000000000987**

離散Bayes識別則の研究成果3：進行大腸がんの再発予測

背景

治癒的切除後の進行がんに対して予後良好の予測は困難で、有効な標的ﾏｰｶｰが求められている。

結果

10ﾏｰｶｰ候補の中から離散Bayes識別則により3標的ﾏｰｶｰを選択し、**ﾃｽﾄｻﾝﾌﾟﾙに対して感度71%, 特異度67%の精度で進行大腸がんの再発を予測できた。**

CD4 and FOXP3 as predictive markers for the recurrence of T3/T4a stage II colorectal cancer: applying a novel discrete Bayes decision rule, *BMC Cancer*, 2022. doi.org/10.1186/s12885-022-10181-7

離散Bayes識別則の研究成果4：抗うつ薬の治療反応性の予測

背景

測定系の影響を受けやすい遺伝子情報を用いて抗うつ薬治療効果を予測することは困難であるが、臨床応用上重要であるため、その確立が求められている。

結果

広島大の訓練ｻﾝﾌﾟﾙを用いて8ﾏｰｶｰ候補の中から離散Bayes識別則により3標的ﾏｰｶｰを選択し、他施設である山口大及び徳島大の**ﾃｽﾄｻﾝﾌﾟﾙに対して感度91%, 特異度75%の精度で治療効果を予測できた。**

Interferon signaling and hypercytokinemia-related gene expression in the blood of antidepressant non-responders, *Heliyon*, 2023. doi: Heliyon, 9(1), art. no. e13059

離散Bayes識別則の研究成果5: 術後合併症の予測（論文投稿中）

背景

術後合併症の予測は、術後管理が大変で医療従事者を疲弊させないために、また患者にとっても重要である。

結果

患者を2群に層別化し、山口大の訓練ｻﾝﾌﾟﾙを用いて群毎に13ﾏｰｶｰ候補の中から離散Bayes識別則により標的ﾏｰｶｰを選択し、他施設である大阪大のﾃｽﾄｻﾝﾌﾟﾙに対して**2群全体として感度86%, 特異度71%の精度で合併症を予測できた。**

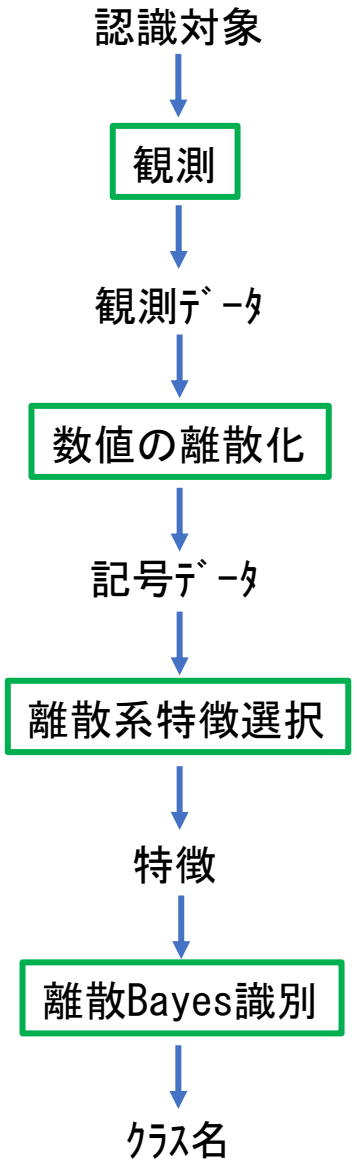
A novel prediction model of pancreatic fistula after pancreaticoduodenectomy using only preoperative markers, to be submitted

基幹となる離散Bayes識別則の特許

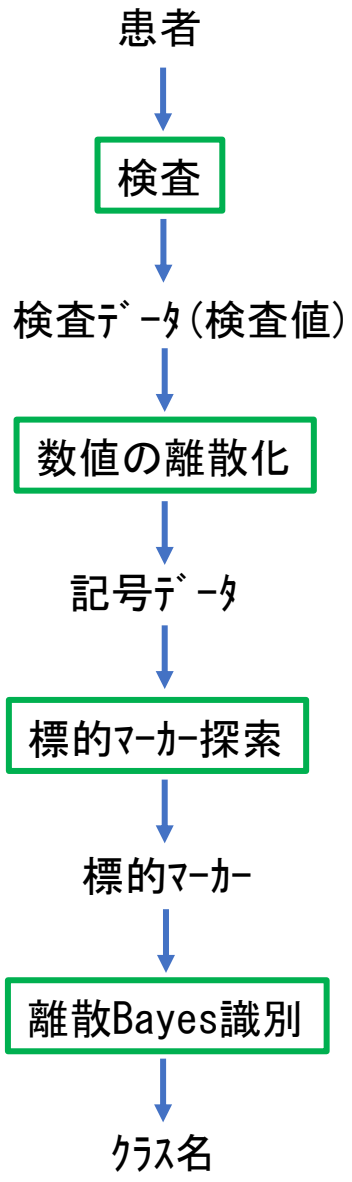
① 特許第6041331号
　出願日 平成28年2月26日（早期審査対象出願） 登録日 平成28年11月18日 特許権者 山口大学
② 特許第6889428号（分割）
　出願日 平成28年2月26日 登録日 令和3年5月25日 特許権者 山口大学
③ Patent No.: US 11,461,598 国際出願日2016年12月26日 特許証発行日2022年10月4日 特許権者 山口大学

離散Bayes識別則による統計的ﾊﾟﾀｰﾝ認識の新展開

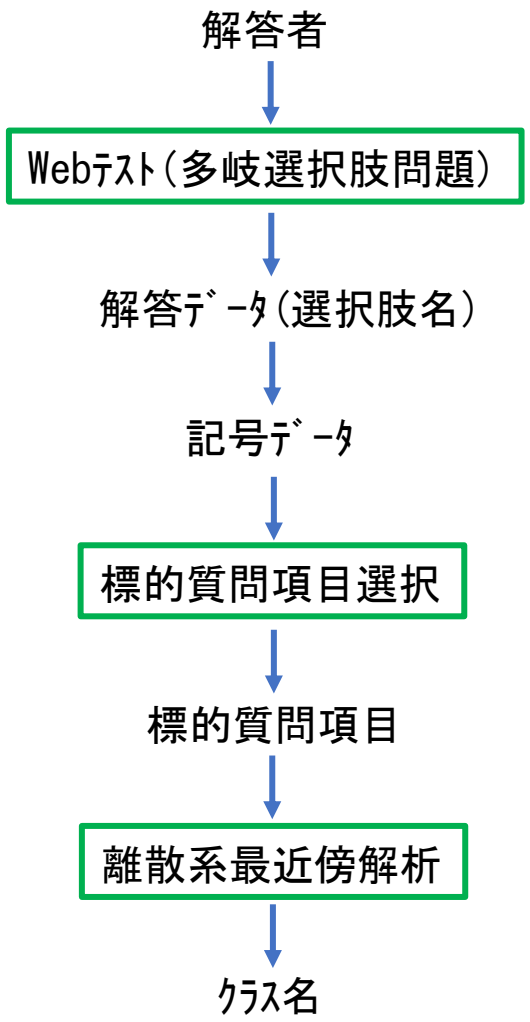
一般論



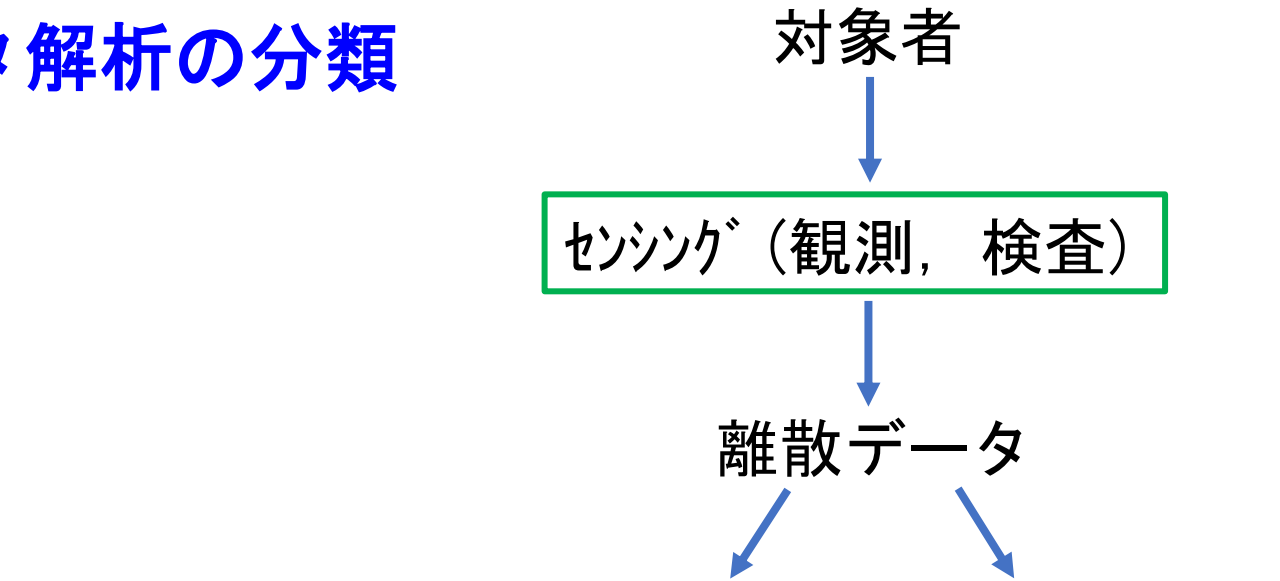
医学分野：①個別化医療，②患者層別化（治験）



Web分野：①eﾗｰﾆﾝｸﾞ，②問診，ｱﾝｹｰﾄ解析



離散データ解析の分類



分類		客観的	主観的
正解	有	検査ﾃﾞｰﾀを用いた診断と予測	ﾃｽﾄの解答ﾃﾞｰﾀを用いた学習支援
	無	病態等の解明	問診やｱﾝｹｰﾄの回答ﾃﾞｰﾀを用いた医療ﾒﾝﾀﾙﾍﾙｽｹｱ

未踏のﾃｰﾏ

無意識下における思い込みや病的なこだわりに対するﾊﾞｲｱｽ解析

まとめ

既存の統計的ﾊﾟﾀｰﾝ認識
（数値ﾃﾞｰﾀのみ対象）
①識別
・Bayes識別則
・最近傍識別則
②特徴選択　（標的ﾏｰｶｰ探索）

(1) 大規模な高速計算機が必要

(2) ﾃﾞｰﾀは改正個情法の対象

優位性

独自の統計的ﾊﾟﾀｰﾝ認識（記号ﾃﾞｰﾀを扱える）
①離散系識別　（患者層別化）
・離散Bayes識別則
・離散系最近傍識別則（Webﾃｽﾄ，ｱﾝｹｰﾄの記号ﾃﾞｰﾀ解析も可）
②離散系特徴選択　（標的ﾏｰｶｰ探索）

(1) 大規模な高速計算機は不要
「次元の呪い」がない。ﾃﾞｰﾀ数が増加しても
計算量の急激な増大はない。
低価格のﾃﾞｨｽｸﾄｯﾌﾟ型PCで解析可能（院内で解析可能）

(2) ﾃﾞｰﾀは改正個情法の対象外
（統計情報処理による非個人情報化）
AI開発で必須である「ﾃﾞｰﾀの第三者提供」に道が開かれる

「こと」化のﾒﾘｯﾄ
AIの結果「こと（医学的概念）」の組合せ（知識の生成）
医師は、医学知識としてAIの結果を理解できる。
これまで単独で用いられていた遺伝子変異が組合せで評価され、新しい可能性が期待される。

参考文献

１．渡辺　慧、認識とﾊﾟﾀﾝ、岩波新書、1978
（若い頃、何回も繰り返して読んだ本です。研究自身には直接関係なかったのですが、
思想的な影響はかなり受けました。）

２．渡辺　慧、知るということ、認知科学選書、東京大学出版会、1986.
（こちらの方が入手し易い　ちくま学芸文庫　ISBN:978-4-480-09381-3）

３．C.R.Rao（藤越、柳井、田栗共訳）、統計学とは何か、丸善、1993.

４．奥野、久米、芳賀、吉澤 共著、多変量解析法、日科技連、1971.

５．浜本義彦、統計的ﾊﾟﾀｰﾝ認識入門、森北出版、2009.

ご清聴、ありがとうございました。

本日の講演原稿のPDFファイルをご希望の方には送付致します。
お問合せは下記までお願い致します。

hamamoto@yamaguchi-u.ac.jp