離散 Bayes 識別則と
その個別化医療への応用に
関する研究

A Study on Discrete Bayes Decision Rule and its Application to Personalized Medicine

平成２９年９月

荻　原　　宏　是

山口大学大学院医学系研究科

学位論文の要旨

癌は生涯で日本人の２人に１人は罹患し、３人に１人はそれが原因で死亡する病気である。この癌の克服は国民的課題であり、一刻も早い治療方法の確立が望まれている。癌治療の困難さは、たとえ癌種が同じであっても患者個々によって癌が異なる多様性にある。現在、癌治療には血液検査やCTなどの多種多様な検査が行われているが、それらは癌の一側面しか診ておらず、単一で決め手となる検査はない。そこで１つの検査だけではなく複数の検査により個々の患者に最適な医療が実現されるように検討が行われている。これは「個別化医療」の考えに基づいている。

個別化医療とは、患者の遺伝情報や現在の疾患の状態を基に患者個々に最適な治療方法を提供する医療である。個別化医療では、患者が病気であると診断されると、患者個々に治療効果が見込め、かつ副作用の少ない治療を提供できる。また、治療効果が薄いと予測される薬の投薬を投与前に防ぐことで、患者の身体的、金銭的負担を抑えることにも繋がるという利点がある。癌の多様性に対処するためには、症状が同じ患者を等しく扱う従来の「平均的医療」よりも患者個々に応じた医療である「個別化医療」が適していると考えられる。

個別化医療に関する研究は、医師が構築したスコア式による予測・診断とコンピュータの機械学習による予測・診断に大別される。まずスコア式による予測・診断については、マーカーからの測定値とそのマーカーに設定したカットオフ値との大小関係でスコア値を付け、スコア値の総和により患者を層別化するアプローチである。しかし、スコア式は医師の経験と直観に基づいて構築さ

れたものであるため、用いられるマーカーに論理的根拠はなく、スコア式の最適性も不明確である。

一方、コンピュータの機械学習による予測・診断は、20 世紀の分子生命科学の成果である遺伝子解析を基に、遺伝子の発現量を用いて白血病の診断が可能となった。これが引き金となり、遺伝子解析による発現データを用いて癌研究が本格的に展開されるようになった。遺伝子関連のデータは数万から数十万とデータ量が膨大であることから、コンピュータの機械学習による診断が注目された。しかし、この機械学習による予測・診断にも深刻な問題点がある。それは、用いるデータが量的データ（数値データ）に限定されている点である。一般に医学データには数値で表される量的データと数値で表すことができない質的データが混在する。量的データと質的データはいずれも癌の予測・診断に本質的であるが、質的データは機械学習の識別器では取り扱うことはできない。機械学習と同じく、統計的パターン認識の Bayes 識別器もまた、質的データには適用できない。

本論文では、質的データを用いることができるように、Bayes 識別則を拡張した「離散 Bayes 識別則」を提案する。診断アルゴリズムに Bayes 識別の考えを採用することで、個々のマーカーの不確実性を数量化して統計的に誤りを最小化できる。さらに、提案手法は、質的データはもちろんのこと、量的データもカットオフ値を設けて 2 値化することにより質的データとし、質的データと量的データの両方を取り扱うことができる。

本論文では、従来の個別化医療の問題点を解決する離散 Bayes 識別則を用いて個別化医療の実現を目的とする。

第 1 章の序論では、個別化医療の必要性と従来の個別化医療の問題点について指摘し、これらを踏まえて本論文の目的と構成を述べる。

第２章では、統計的パターン認識を概説し、従来の個別化医療問題を統計的パターン認識問題として定式化する。その後、離散 Bayes 識別則を提案し、その評価指標も説明する。

第３章では、離散 Bayes 識別則を、肝癌の早期再発の予測問題、早期胃癌のリンパ節転移の予測問題、大腸癌における抗癌剤と免疫療法の併用効果の予測問題、漢方薬の処方問題に適用し、提案手法の有用性を検討する。

第４章では、結論として本論文の総括と今後の展望について述べる。

Abstract

Cancer has a lifetime affecting one out of two Japanese people, one in three people dying of it due to it. Overcoming this cancer is a national subject, and establishment of a treatment method as soon as possible is desired. The difficulty of treating cancer due to diversity of cancer, even if the cancer types are the same, but depending on individual patients. The present cancer therapy is undergoing a wide variety of tests such as blood tests and CT, but they have examined only one aspect of cancer. Unfortunately, there is no effective test. Therefore, multi-dimensional data produced by use of various tests is used to represent the medical conditions of a patient and using the data, personalized medicine is conducted.

Personalized medicine is medical treatment that selects the optimal treatment method for each patient based on the genetic information of the patient and the state of the current disease. In personalized medicine, if a patient is diagnosed as having a disease, it is possible to select a treatment with a therapeutic effect expected for each patient and few side effects. In addition, by preventing medication that is expected to have a low therapeutic effect, there is an advantage that it also leads to suppressing the physical and financial burden of the patient. In order to cope with the diversity of cancer, it is considered that "personalized medicine", which is the optimal medical care for each patient, is more suitable than "conventional medicine" which treats patients with the same symptoms equally.

Research on personalized medicine can be roughly divided into the score

formula approach established by a medical doctor and machine learning approach. In the score formula approach, a score value is produced according to the relationship between the measurement value from the marker and the cutoff value set for the marker, and the state of the patient is stratified by the sum of the score values. However, since the score formula approach is constructed based on the experience and intuition of a medical doctor, there is no theoretical basis for the marker used, and the optimality of the score formula approach is also unclear.

Meanwhile, machine learning approach makes it possible to diagnose leukemia using the expression level of genes based on gene analysis which is the result of molecular life science of the 20th century. This triggered the development of cancer research in full scale using expression data by microarray analysis. Diagnosis by machine learning has attracted attention, because the data amount of genetic related data is huge from tens of thousands to hundreds of thousands. However, this machine learning approach also has a serious problem that the data used is limited to quantitative data (numerical data). Though both quantitative data and qualitative data in medical data are essential for personalized medicine of cancer, qualitative data can not be handled by machine learning. Like machine learning, the Bayes classifier in statistical pattern recognition can not be also applied to qualitative data.

In this paper, "discrete Bayes decision rule" is proposed. The advantage of the proposed method is that it can deal with qualitative data. The Bayes approach that quantifies the uncertainty of each marker and minimizes

statistically errors due to the uncertainly is adopted. Furthermore, quantitative data are transformed to qualitative data by binarizing of setting an optimal cutoff value, and the proposed method can handle both qualitative data and quantitative data.

The purpose of the paper is to realize personalized medicine by the discrete Bayes decision rule which overcomes the problem of conventional machine learning approach for personalized medicine.

In Chapter 1, the necessity of personalized medicine and the problem of conventional personalized medicine are pointed out. Next, the purpose and composition of this thesis are described.

In Chapter 2, statistical pattern recognition is outlined and the personalized medical problem as one of statistical pattern recognition problems is formulated. After that, a novel discrete Bayes decision rule is proposed and the evaluation methods are described.

In Chapter 3, the discrete Bayes decision rule is applied to the problem of predicting early recurrence in liver cancer, the problem of predicting lymph node metastasis in early gastric cancer, the problem of predicting the effectiveness of combination of anticancer drug and immunotherapy in colon cancer, the problem of prescribing Kampo medicine, and the usefulness of the proposed method for each medical problem is discussed.

In Chapter 4, this paper is summarized. Based on the summary, conclusions and the future prospects of the proposed method are described.

目次

第１章　序論

1.1 はじめに

癌は生涯で日本人の２人に１人は罹患し、３人に１人はそれが原因で死亡する病気である[1]。癌の克服は国民的課題であり、内閣府の政策の１つである総合科学技術・イノベーション会議において継続的に取組む目標とされている[2]。癌治療の困難さは、たとえ癌種が同じであっても患者個々によって癌が異なる多様性にある。しかも、血液検査や CT などの多種多様な検査も癌の一側面しか診ておらず、単一で決め手となる検査はない。そこで１つの検査だけではなく複数の検査から得られるデータを用いて個々の癌を多角的に見るのが常である。それゆえ高精度な診断を実現するために個々の患者に応じた「個別化医療」[3]という考え方が生まれた。

この個別化医療の必要性を説くために、まず従来の医療の問題点を指摘する。従来の医療では、例えば診療情報から患者が病気であると診断されると、その病名に応じた薬を処方する。この処方薬は患者の体質はまったく考慮されないため、薬が有効であるか、副作用が出るか否かは投薬後にしか分からない。また、このような体質の個人差は、効果が表れるまでの期間や効果の持続時間など随時観察しなければ分からないものであり、患者が効果を実感するに至るまでには紆余曲折がある。

一方、「個別化医療」は患者個々に応じた医療であり、患者の遺伝情報や現在の疾患の状態を基に患者個々に最適な治療方法を選択することを目的とする。個別化医療のイメージを図 1.1 として示す。

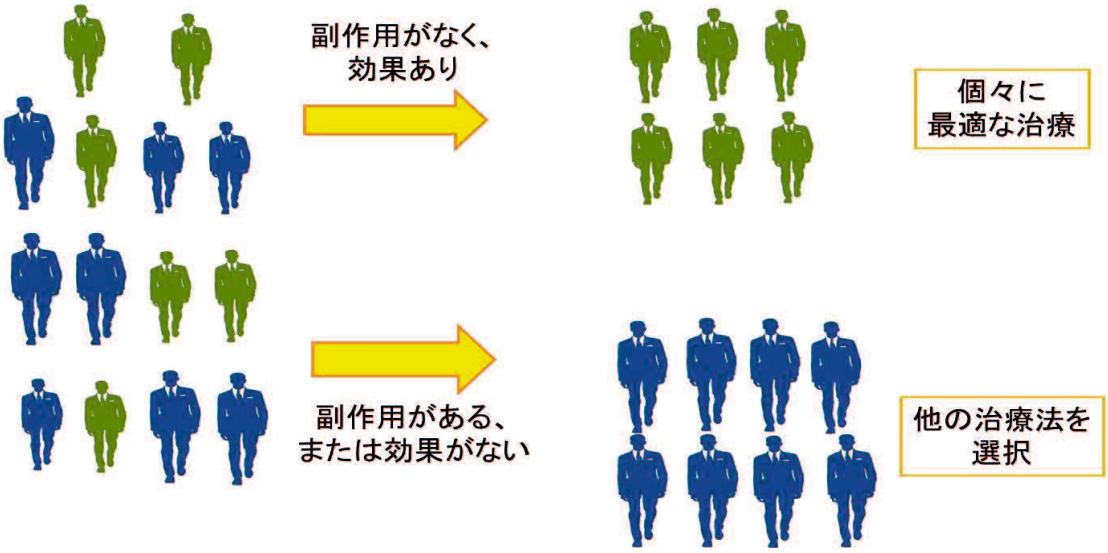


図 1.1 個別化医療のイメージ

個別化医療では、患者が病気であると診断されると、まずコンパニオン診断（Companion diagnostics；CoDx）により、どの医薬品が有効か、また副作用の有無は？等の観点から患者を識別（層別化）する。そして、効果があり、副作用の少ない患者グループに対して分子標的薬の投薬治療をする。それ以外の患者グループは他の治療方法を選択できる。この個別化医療では患者個々に治療効果が見込め、かつ副作用の少ない治療を選択できる。また、治療効果が低いと予測される薬の投与を防ぐことで患者の身体的、金銭的負担を抑えることにも繋がるという利点がある。その結果、医療費の高騰を抑制できる。癌の多様性に対処するためには、症状が同じ患者を等しく扱う従来の「平均的医療」よりも患者個々に応じた医療である「個別化医療」が適していると考えられている。

1.2 従来の個別化医療に関する研究

従来の個別化医療に関する研究は、医学的知見に基づいて医師によって構築されたスコア式による予測・診断とコンピュータの機械学習による予測・診断に大別される。以下、代表的な研究を紹介する。

1.2.1 医師によるスコア式を用いた予測・診断

第３章で取り扱う肝癌を例に説明する。従来は、肝臓の状態をスコア式として表わし、医師はスコア式の値の高低で診断・予測などを行ってきた。これまでスコア式は数多く提案されているが、その中でも代表的なスコア式を３つ説明する。

Tokyo Score[4]は、東京大学で作成された肝臓のスコア式である。定義を表 1.1 に示す。スコア式に用いられるマーカー（検査、特徴）として、アルブミン、ビリルビン、腫瘍サイズ、腫瘍数の４つがある。それぞれのマーカーには０～２点のスコア値として配点があり、それらの総和である総スコア値は０～８点となる。一般に肝臓の状態が悪い程、総スコア値は高くなる。

次に Modified JIS[5]は、本邦で提案されたスコア式で、肝癌患者の生存率をより正確に予測するために The Japan Integrated Staging score (JIS score)[6]を修正したものである。定義を表 1.2 に示す。スコア式に用いられるマーカーは、Liver damage（肝障害度）と TNM 分類[7],[8]である。いずれのマーカーも肝臓の状態を表すスコア式である。Liver damage に 0~2 点、TNM 分類に 0~3 点のスコア値が配点され、総スコア値は 0~5 点となる。肝臓の状態が悪い程、総スコア値の点数は高くなる。

表 1.1　Tokyo Score の定義

マーカー	点数		
	0	1	2
アルブミン（g/dL）	>3.5	2.8-3.5	<2.8
ビリルビン（g/dL）	<1	1-2	>2
腫瘍サイズ（cm）	<2	2-5	>5
腫瘍数	<3	3	>3

表 1.2　Modified JIS の定義

マーカー	点数			
	0	1	2	3
Liver damage	A	B	C	
TNM分類	Ⅰ	Ⅱ	Ⅲ	Ⅳ

最後に、TNM 分類は、肝癌取扱い規約におけるステージングで、よく知られている。定義を表 1.3 に示す。スコア式に用いられるマーカーとして、T 因子（腫瘍数、腫瘍サイズ、脈管侵襲）、N 因子（リンパ節転移）、M 因子（遠隔転移）がある。TNM 分類ではマーカー毎のスコア値の総和をとる他のスコア式とは異なり、マーカー毎にその患者の進行度を求め、もっとも高い数値をその患者の進行度とする。なお、この TNM 分類は肝癌に限定されたものではなく、様々な癌種に応じたスコア式がある。

表 1.3　肝癌に対する TNM 分類の定義

因子 / stage	T因子	N因子	M因子
Ⅰ	T1	N0	M0
Ⅱ	T2	N0	M0
Ⅲ	T3	N0	M0
ⅣA	T4	N0	M0
	T1, T2, T3, T4	N1	M0
ⅣB	T1, T2, T3, T4	N0, N1	M1

T因子	T1	T2	T3	T4
①腫瘍個数 単発 ②腫瘍径 2cm以下 ③脈管侵襲なし	3項目合致	2項目合致	1項目合致	全て合致せず

N因子	N0	N1
リンパ節転移	なし	あり

M因子	M0	M1
遠隔転移	なし	あり

これらスコア式による予測・診断には、共通して、診断アルゴリズムがヒューリスティック（発見的）であるという問題点がある。従来、スコア式には単一マーカーが用いられることが多く、カットオフ値の最適化問題があるものの、マーカーが陽性か陰性かの判断だけで診断・予測は容易であった。しかし、決め手となる決定的なマーカーが見つからない場合、複数のマーカーを用いることになり、これをマーカーの多層化と呼ぶが、このとき不確実な多数のマーカーを用いて如何に診断・予測するかについては、十分な検討がなされていない。スコア式が医師の経験と直観に基づいて構築されたものであるがゆえに、マーカーまたスコア式自身に論理的根拠がない。

1.2.2 機械学習による予測・診断

従来の機械学習による予測・診断では、教師あり学習が用いられている[18]。教師あり学習とは、正解のクラスラベルが付与されているサンプルを用いるコンピュータの学習である。教師あり学習による機械学習を用いて、これまで様々な癌の分類が報告されている。代表的な研究をいくつか紹介し、それらを表 1.4 にまとめる。

Khan ら[9]により、人工ニューラルネットワーク（ANN）を用いた癌の分類が報告されている。４種の異なる small round blue cell tumors (SRBCTs)は臨床診療において診断が難しいとされるが、ANN を用いた診断で全てのサンプルを正しく分類し、分類に最も関連する遺伝子を同定した。また、識別性能を調べるために、訓練サンプル以外のサンプルを用いてテストを行った結果、全てのケースで正しく分類した。

Furey ら[10]により、サポートベクターマシン（SVM）を用いた癌の分類が報告されている。97802 個の cDNA の発現データセットを用いて卵巣癌組織と正常卵巣組織の２クラスでの識別を行った。その結果、癌組織を完全に分類することに成功した。

Iizuka ら[11]により、肝癌の再発予測システムが報告されている。約 7000 遺伝子を特徴の候補として肝癌患者を含む 33 名の訓練サンプルに対し 12 遺伝子を特徴として選択し、Fisher 線形識別器を用いて予測システムを構築した。この予測システムの識別性能を、訓練に用いた症例とは別の 27 名のテストサンプルで識別し、識別率 93%を達成した。一方、SVM の識別率は 60%であり、提案手法が SVM の識別性能を上回った。

表 1.4　従来の機械学習による予測・診断

文献番号	筆頭著者	癌種	診断内容	クラス	マーカー	識別器	年
[9]	J. Khan	small round blue cell tumors (SRBCTs)	癌の分類	neuroblastoma (NB) rhabdomyosarcoma (RMS) non-Hodgkin lymphoma (NHL) Ewing family of tumors (EWS)	cDNAマイクロアレイ	ANN	2001
[10]	T.S. Furey	卵巣癌	癌の分類	卵巣癌組織 正常卵巣組織 他の正常組織	cDNAマイクロアレイ	SVM	2000
[11]	N. Iizuka	肝癌	癌の再発予測	肝癌1年以内再発 無再発	高密度オリゴヌクレオチド マイクロアレイ	Fisher線形識別器	2003
[12]	R. Tibshirani	small round blue cell tumors (SRBCTs)	癌の分類	Burkitt lymphoma (BL) Ewing sarcoma (EWS) neuroblastoma (NB) rhabdomyosarcoma (RMS)	cDNAマイクロアレイ	最近シュランケン重心法	2002
[13]	M. Xiong	大腸癌、乳癌	癌の分類	大腸腫瘍組織 乳房腫瘍組織 急性リンパ芽球性白血病 急性骨髄性白血病	マイクロアレイ	Fisher線形識別器	2001
[14]	I. Guyon	大腸癌	癌の分類	大腸腫瘍組織 正常組織	DNAマイクロアレイ	SVM	2002
[15]	K. Kourou	肺癌、乳癌、口腔癌等	癌の予後予測	癌組織 正常組織	遺伝子発現 臨床検査データ等	ANN、SVM、 Decision Treeなど	2015
[16]	N. Iizuka	肝癌	癌の診断	肝癌患者 健常者	メチル化遺伝子 血液検査データ	Fisher線形識別器	2011
[17]	L. Parthiban	乳癌	癌の診断	良性腫瘍 悪性腫瘍	形状など腫瘍に関する情報	Coactive Neuro-Fuzzy Inference System (CANFIS)	2009

Tibshirani ら[12]により、small round blue cell tumors (SRBCTs)及び白血病の分類が報告されている。クラスター分析の一つである重心法を基に提案された「最近シュランケン重心法」により、small round blue cell tumors (SRBCTs)及び白血病を分類するための遺伝子を同定した。

Xiong ら[13]により、分子的な腫瘍分類について報告されている。データセットには、22 の正常及び 40 の大腸腫瘍組織における 2000 遺伝子、14 の乳房上皮細胞及び 13 の乳房腫瘍における 5776 の配列、47 個の急性リンパ芽球性白血病及び 25 個の急性骨髄性白血病における 6817 個の遺伝子の発現データを用いた。2 つまたは 3 つの遺伝子の組合せによる Fisher 線形識別器を用いた識別により、識別率 90%を超えた。

Guyon ら[14]により、DNA マイクロアレイによる癌の分類が報告されている。再帰的特徴除去（RFE）に基づくサポートベクターマシン（SVM）による識別

を提案した。白血病の患者に対して leave-one-out 法による最高性能には 64 個の遺伝子を必要とした。大腸癌の患者に対しては、４つの遺伝子を用いて 98%の識別率を達成した。

Kourou ら[15]により、機械学習を用いた癌分類が報告されている。機械学習を用いた例として、人工ニューラルネットワーク（ANN）、ベイジアンネットワーク(BN)、サポートベクターマシン（SVM）、決定木（DT）など、様々な識別器が癌分類に適用された。

Iizuka ら[16]により、メチル化量を特徴に用いた血液検査による肝癌診断が報告されている。これまで C 型肝炎ウイルス（HCV）感染に関連する肝細胞癌（HCC）を効率的に検出するための血液検査はほとんどないが、108 人の HCC 患者と 56 人の健常者のデータセットに対して AFP、PIVKA‐Ⅱと２つのメチル化遺伝子（SPINT2 および SRD5A2）という４マーカーを用いた識別器により、識別率 82.3%という高い識別性能を示した。この識別器の性能を４つの他施設からのデータを合わせた 112 人の HCC 患者と 146 人の健常者のデータで評価し、識別率 81.4%を達成した。

Parthiban ら[17]により、ファジイ推論システムとニューラルネットワークの両方の利点を持つ CANFIS を用いた乳癌の診断システムが報告されている。この CANFIS を用いた識別手法は、SVM や他のニューラルネットワークの手法と比べて最も高い識別性能（PPV）を示した。

これら機械学習による予測・診断には、従来の識別器はいずれも量的データ（数値データ）しか利用できない問題点がある。一般に医学データには数値で表される量的データと数値で表すことができない質的データ（記号データ）が混在する。例えば量的データとして人間ドッグの GOT や GPT などがあり、質的データとしては遺伝子変異の有無や検査の陽性・陰性などの記号がある。こ

れら質的データは、癌の予測・診断に本質的なマーカーであるが、機械学習では取り扱うことはできない。上述の機械学習と同じく、統計的パターン認識[18]の Bayes 識別器もまた、質的データには適用できない。

1.3 本論文の目的と構成

本論文では質的データを用いることができないという従来の機械学習の問題点を解決する手法として、Bayes 識別則を拡張した「離散 Bayes 識別則」を提案する。診断アルゴリズムに Bayes 識別の考えを採用することで、マーカーの不確実性を数量化して理論上誤りを最小化できる。さらに、提案手法は、質的データはもちろんのこと、量的データもカットオフ値により 2 値化し、質的データとすることで、質的データと量的データが混在していても取り扱うことができる。本論文では、従来の問題点を解決する離散 Bayes 識別則を用いた個別化医療の実現を目的とする。

第 1 章の序論では、個別化医療の必要性と従来の個別化医療の問題点について指摘し、これらを踏まえて本論文の目的と構成を述べる。

第 2 章では、統計的パターン認識を概説し、従来の個別化医療問題を統計的パターン認識問題として定式化する。その後、離散 Bayes 識別則を提案し、その評価方法についても述べる。

第 3 章では、離散 Bayes 識別則を、肝癌の早期再発の予測問題、早期胃癌のリンパ節転移の予測問題、大腸癌における抗癌剤と免疫療法の併用効果の予測問題、漢方薬の処方問題に適用し、提案手法の有用性を検討する。

第 4 章では、結論として本論文の総括と今後の展望について述べる。

1.4 準備

本論文で用いる記号について説明する。

x：パターンベクトル

M：マーカー候補数

N：独立試行回数

d：標的マーカー数

n：サンプル数（患者数）

m：クラス数

ω_i：クラスi

r_j：分割に対する番号

$x_{j(r_j)}$：マーカーx_jのr_j番目の分割を$x_{j(r_j)}$

$n^i_{j(r_j)}$：クラスω_iの患者n^i人の中で、分割$x_{j(r_j)}$に属する患者数

$P(x_{j(r_j)}|\omega_i)$：クラスω_iにおける分割$x_{j(r_j)}$の条件付き確率

$P(\omega_i)$：クラスω_iの事前確率

$P(\omega_i|\mathbf{x})$：パターン x に対するクラスω_iの事後確率

$P(y,\mathbf{x})$ ：同時確率（漢方薬 y とパターン $\mathbf{x}$ が同時に起こる確率）

参考文献

1)最新がん統計：国立がん研究センター　がん登録・統計, http://ganjoho.jp/reg_stat/statistics/stat/summary.html.

2)総合科学技術会議（第 111 回）議事次第 ー総合科学技術・イノベーション会議ー内閣府, http://www8.cao.go.jp/cstp/siryo/haihu111/siryo1-1.pdf .

3) 医薬品の開発・承認審査に関わる個別化医療の現状評価に関する議論の取りまとめ 資料 1-2, https://www.pmda.go.jp/files/000155940.pdf.

4) R. Tateishi, H. Yoshida, S. Shiina et al., Proposal of a New Prognostic Model for Hepatocellular Carcinoma: an Analysis of 403 Patients, Gut, Vol.54, pp.419-425, 2005.

5) H. Ikai, K. Takayasu, M. Omata et al., A Modified Japan Integrated Stage Score for prognostic assessment in patients with hepatocellular carcinoma, Journal of Gastroenterology, Vol.41, pp.884-892, 2006.

6) M. Kudo, H. Chung, S. Haji, Y. Osaki, H. Oka, T. Seki, H. Kasugai, Y. Sasaki, T. Matsunaga, Validaton of a new prognostic staging system for hepatocellular carcinoma: the JIS score compared with the CLIP score, Hepatology, Vol.40, No.6, pp.1396-1405, 2004.

7) M. Minagawa, I. Ikai, Y. Matsuyama, Y. Yamaoka, and M. Makuuchi, Staging of hepatocellular carcinoma: Assessment of the Japanese TNM and AJCC/UICC TNM systems in a cohort of 13,772 patients in Japan, Annals of Surgery, Vol.245, No.6, pp.909-922, 2007.

8) J.M. Henderson, M. Sherman, A. Tavill, M. Abecassis, G. Chejfec, and T. Gramlich, AHPBA/AJCC Consensus Conference on Staging of Hepatocellular Carcinoma: Consensus Statement, HPB (Oxford), Vol.5, No.4, pp.243-250, 2003.

9) J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold,

M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Medicine, Vol.7, pp.673– 679, 2001.

10) T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer and D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics, Vol.16, No.10, pp.906-914, 2000.

11) N. Iizuka, M. Oka, Y. Hamamoto et al., Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, Lancet, Vol.361, pp.923-929, 2003.

12) R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proceedings of the National Academy of Sciences of the United States of America, Vol.99, pp.6567– 6572, 2002.

13) M. Xiong, W. Li, J. Zhao, L. Jin, E. Boerwinkle, Feature (gene) selection in gene expression-based tumor classification, Molecular Genetics and Metabolism, Vol.73, pp.239– 247, 2001.

14) I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning, Vol.46, pp.389– 422, 2002.

15) K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, Vol.13, pp.8-17, 2015.

16) N. Iizuka, M. Oka, I. Sakaida, T. Moribe, T. Miura, N. Kimura, S. Tamatsukuri, H. Ishitsuka, K. Uchida, S. Terai, S. Yamashita, K. Okita, K. Sakata, Y. Karino, J. Toyota, E. Ando, T. Ide, M. Sata, R. Tsunedomi, M. Tsutsui, M. Iida, Y. Tokuhisa, K. Sakamoto, T,

Tamesa, Y. Fujita, Y. Hamamoto, Efficient detection of hepatocellular carcinoma by hybrid blood test of epigenetic and classical protein markers, Clinica Chimica Acta, Vol.412, pp.152-158, 2011.

17) L. Parthiban, R. Subramanian, CANFIS—a computer aided diagnostic tool for cancer detection, Journal of Biomedical Science and Engineering, Vol.2, pp.323-335, 2009.

18) A.K. Jain, R.W. Duin, and J. Mao, Statistical pattern recognition: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.1, pp.4-37, 2000.

第２章　離散 Bayes 識別則

2.1 統計的パターン認識

統計的パターン認識[1]とは、認識対象をパターンとして表し、それらパターンのなす分布に着目し、その統計的構造に関する知識からパターン認識問題を解くものである。ここでは、この統計的パターン認識を説明する。

図 2.1 にパターン認識モデルを示す。パターン認識の過程は、認識対象の観測、特徴の選択、識別の３つの処理に大別される。

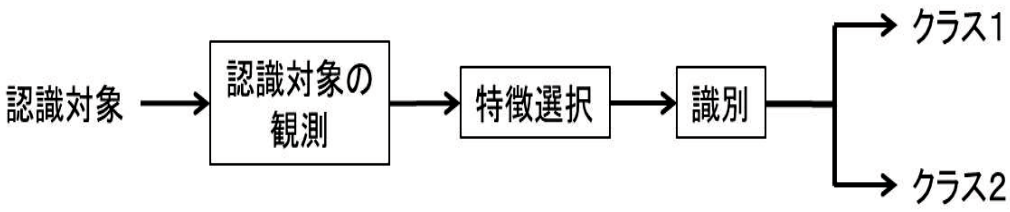


図 2.1 パターン認識モデル

初めに、外界に存在する認識対象を観測することで認識対象を観測データの組として表わす。次に、その得られた観測データの中から特徴選択により識別に有用な特徴を取り出し、得られた最適な特徴の組を用いて識別器を設計し、認識対象がどのクラスに属するかを識別する。

ここで、パターン認識の処理の中で識別以外の、観測、特徴選択はいずれも認識対象の性質による影響が大きい。よって、認識対象の性質を処理に反映した手法が用いられる。しかし、認識対象の性質は未だ解明されていないものも多い。このような問題に対してパターンのなす統計的構造に着目し、それに基づいてパターン認識問題を解く理論が統計的パターン認識の理論である。

2.2 個別化医療問題のパターン認識問題としての定式化

前述したように個別化医療では患者によって治療が異なる。これは、患者を入力すれば、どの治療を行うべきかというパターン認識の問題とみなせる。この個別化医療問題を一般論として統計的パターン認識問題として表現すると、以下の図 2.2 で表される。患者に対してマーカーを測定することによりマーカーの測定値を得る。次に、マーカーの中からマーカー選択により識別に有用なものを最適な標的マーカーとして選択し、得られた最適な標的マーカーを用いて識別器を設計して、患者がどのクラスに属するかを識別する。

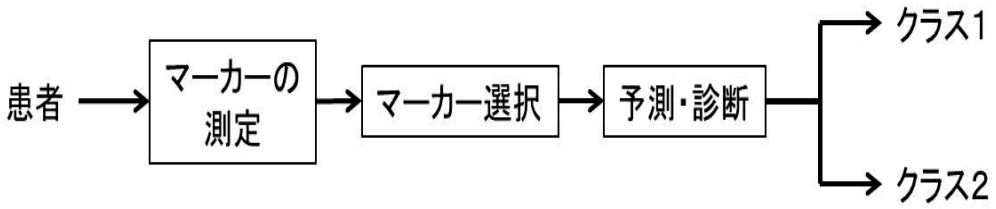


図 2.2　個別化医療問題に適用した場合のパターン認識モデル

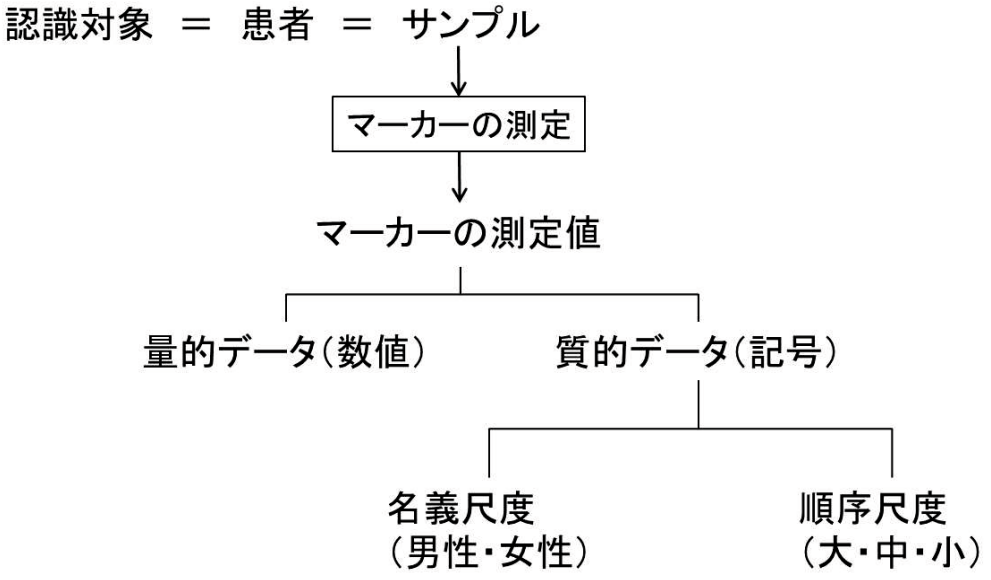


図 2.3 用語の説明

本論文で取り扱う個別化医療問題を統計的パターン認識問題として定式化すると、以下の表 2.1 で表わされる。認識問題におけるクラスは、肝癌の再発予測であれば再発の有無、胃癌のリンパ節転移の予測であれば転移の有無、大腸癌の抗癌剤と免疫療法の併用効果の予測であれば併用効果の有無、漢方薬の処方であれば漢方薬となる。漢方薬の処方ではクラスが漢方薬であるとは、入力となる症状を訴える患者にある漢方薬を対応付ける（漢方薬を処方する）という意味を持つ。いずれの問題においても認識対象は患者で、マーカーの測定値は質的データ、量的データから構成される。例えば本論文では量的データとしてマイクロ RNA のデータ、質的データとしては症状のデータなどがある。質的データについては、男性・女性のような名義尺度と検査値の大・中・小などの順序尺度に分けられる。以上を整理すると図 2.3 となる。前述したように、これらの質的データは診断・予測に必要とされるが、従来の識別手法では取り扱うことはできない。

表 2.1　個別化医療問題のパターン認識問題としての定式化

医学問題	認識対象	マーカーの測定値（データ）	クラス
肝癌の早期再発の予測	患者	質的、量的データ （臨床データ）	肝癌再発の有無
早期胃癌のリンパ節転移の予測	患者	質的、量的データ （臨床データ）	リンパ節転移の有無
大腸癌の抗癌剤と免疫療法の 併用効果の予測	患者	量的データ （遺伝子情報：マイクロRNA）	併用効果の有無
漢方薬の処方	患者	質的データ （症状）	漢方薬

2.3 離散 Bayes 識別則

離散 Bayes 識別器は、通常の Bayes 識別器と異なり、質的データを取り扱うことができる点に特長がある。M 個のマーカー候補が与えられたとき、マーカー選択により M 個の中から識別に有用な d 個のマーカーが選択されたとする。d 個の各マーカーはそれぞれの範囲を互いに排反事象となるように分割すると、患者の測定値はマーカー毎にいずれかの分割に属することになる。いま d 個のマーカー$x_1, x_2, \ldots, x_d$に対して、ある患者の測定値が $x_{1(r_1)}, x_{2(r_2)}, x_{d(r_d)}$の分割に属したとする。ここでマーカー$x_j$の$r_j$番目の分割を$x_{j(r_j)}$と表わす。添字$j$はマーカーの識別番号であり、$j \in \{1,2,\ldots,d\}$である。このとき患者はパターン $\mathbf{x} = [x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)}]^T$と表わされる．分割$x_{j(r_j)}$に入る割合$\mathrm{P}(x_{j(r_j)}|\omega_i)$を

$$\mathrm{P}\left(x_{j(r_j)}\middle|\omega_i\right) = \frac{n^i_{j(r_j)}}{\sum_{k=1}^{d} n^i_{k(r_k)}} \qquad j = 1,2,\ldots,d \tag{1}$$

で定義する．ここで$n^i_{j(r_j)}$はクラスω_iの患者n^i人の中で、分割$x_{j(r_j)}$に属する患者数を表す。このとき、制約条件として、

$$\sum_{j=1}^{d} \mathrm{P}\left(x_{j(r_j)}\middle|\omega_i\right) = 1$$

を設定する。これにより、直観的には各分割に入るサンプル数の割合を条件付き確率とみなすことになる。

一般に、マーカーの測定値が各マーカーのいずれかの分割に入るという事象が互いに独立であると仮定すると、

$$\begin{aligned}
&\mathrm{P}(\mathbf{x}|\omega_i) \\
&= \mathrm{P}(x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)}|\omega_i) \\
&= \mathrm{P}(x_{1(r_1)}|\omega_i)\,\mathrm{P}\left(x_{2(r_2)}\middle|\omega_i\right)\ldots\mathrm{P}\left(x_{d(r_d)}\middle|\omega_i\right) \\
&= \prod_{k=1}^{d} \mathrm{P}(x_{k(r_k)}|\omega_i)
\end{aligned} \tag{2}$$

となる．２クラス問題における事後確率$\mathrm{P}(\omega_i|\,\mathbf{x})$は、Bayesの定理により

$$\mathrm{P}(\omega_i|\,\mathbf{x}) = \frac{P(\omega_i)P(\,\mathbf{x}\,|\omega_i)}{P(\omega_1)P(\,\mathbf{x}\,|\omega_1)+ P(\omega_2)P(\,\mathbf{x}\,|\omega_2)} \tag{3}$$

となる。ここで、$\mathrm{P}(\omega_1)$は事前確率であり、本論文では再発クラスと無再発クラスを等しく扱うために、事前確率$\mathrm{P}(\omega_i)$を等確率の 0.5 として

$$\mathrm{P}(\omega_i|\,\mathbf{x}) = \frac{P(\,\mathbf{x}\,|\omega_i)}{P(\,\mathbf{x}\,|\omega_1) + \ P(\,\mathbf{x}\,|\omega_2)}$$

を得る。これに式（２）を代入すると，事後確率は

$$\mathrm{P}(\omega_i|\,\mathbf{x}) = \frac{\prod_{k=1}^{d}\mathrm{P}(x_{k(r_k)}|\omega_i)}{\prod_{k=1}^{d}\mathrm{P}(x_{k(r_k)}|\omega_1)+\prod_{s=1}^{d}\mathrm{P}(x_{s(r_s)}|\omega_2)} \tag{4}$$

となる。

離散 Bayes 識別則では、パターン$\mathbf{x}$を事後確率$\mathrm{P}(\omega_i|\,\mathbf{x})$が最大のクラス$\omega_i$へ識別する．分割と患者数との関係の一例を表 2.2 に示す。

表 2.2　各マーカーにおける分割と患者数

マーカーの分割	ω_1	ω_2
$x_{1(1)}$	$n^1_{1(1)}$	$n^2_{1(1)}$
$x_{1(2)}$	$n^1_{1(2)}$	$n^2_{1(2)}$
$x_{2(1)}$	$n^1_{2(1)}$	$n^2_{2(1)}$
$x_{2(2)}$	$n^1_{2(2)}$	$n^2_{2(2)}$
$x_{2(3)}$	$n^1_{2(3)}$	$n^2_{2(3)}$
・	・	・
・	・	・
・	・	・
$x_{j(r_j)}$	$n^1_{j(r_j)}$	$n^2_{j(r_j)}$
・	・	・
・	・	・
・	・	・
$x_{d(r_d)}$	$n^1_{d(r_d)}$	$n^2_{d(r_d)}$
・	・	・
・	・	・
・	・	・

表 2.2 ではマーカーx_2の範囲は、$x_{2(1)}$, $x_{2(2)}$, $x_{2(3)}$ の３つに重なりなく分割され、このうち分割$x_{2(2)}$にはクラスω_iの患者n^i人の内$n^i_{2(2)}$人が属していることを示している。ここで$n^i_{2(1)} + n^i_{2(2)} + n^i_{2(3)} = n^i$である。

具体的に例を用いて離散 Bayes 識別則を説明する。いまマーカーx_1とx_2が対象となって、ある患者の測定値が分割$x_{1(1)}$と$x_{2(3)}$に属したとする。

このとき、

$$\mathrm{P}\left(x_{1(1)}\middle|\omega_1\right) = \frac{n^1_{1(1)}}{n^1_{1(1)} + n^1_{2(3)}}$$

$$\mathrm{P}\left(x_{2(3)}\middle|\omega_1\right)=\frac{n_{2(3)}^1}{n_{1(1)}^1+n_{2(3)}^1}$$

であり、$\mathrm{P}\left(x_{1(1)},x_{2(3)}\middle|\omega_1\right)$は

$$\mathrm{P}\left(x_{1(1)},x_{2(3)}\middle|\omega_1\right)=\mathrm{P}\left(x_{1(1)}\middle|\omega_1\right)\mathrm{P}\left(x_{2(3)}\middle|\omega_1\right)$$

となる。同様にして$\mathrm{P}\left(x_{1(1)},x_{2(2)}\middle|\omega_2\right)$を求め、式（４）によりクラス$\omega_1$と$\omega_2$の事後確率をそれぞれ求めて、事後確率が最大のクラスへ患者を識別する。

以下、表 2.2 に実際に数値を入れて事後確率の値を求めてみる。表 2.3 に、マーカーx_1, x_2の具体例を用いて癌の再発予測の流れを示す。

表 2.3　各マーカーにおける分割と患者数

マーカー	マーカーの離散化	再発クラスω_1	無再発クラスω_2
性差	男	30人	20人
	女	10人	80人
腫瘍の大きさ	大	20人	20人
	中	16人	10人
	小	4 人	70人

患者に対するマーカーx_1（性差）, x_2（腫瘍の大きさ）の測定値がそれぞれ分割$x_{1(1)}$（性差は男）と$x_{2(3)}$（腫瘍の大きさは小）に属すると、クラスω_1（再発クラス）の分割に対する条件付き確率は、

$$\mathrm{P}\left(x_{1(1)}\middle|\omega_1\right)=\frac{30}{30+4}$$

$$\mathrm{P}\left(x_{2(3)}\middle|\omega_1\right)=\frac{4}{30+4}$$

である。式（２）より、クラスω_1（再発クラス）に対する条件付き確率は、

$$\mathrm{P}\left(x_{1(1)}, x_{2(3)} \middle| \omega_1\right) = \frac{30}{34} \times \frac{4}{34}$$

$$= 0.104$$

となる。同様に、クラスω_2（無再発クラス）の分割に対する条件付き確率は、

$$\mathrm{P}\left(x_{1(1)} \middle| \omega_2\right) = \frac{20}{20+70}$$

$$\mathrm{P}\left(x_{2(3)} \middle| \omega_2\right) = \frac{70}{20+70}$$

となる。クラスω_2（無再発クラス）に対する条件付き確率は

$$\mathrm{P}\left(x_{1(1)}, x_{2(3)} \middle| \omega_2\right) = \frac{20}{90} \times \frac{70}{90}$$

$$= 0.173$$

となる。これらを式(4)へ代入すると各クラスに対するパターンの事後確率を以下のように計算できる。

$$\mathrm{P}(\omega_1 | \mathbf{x}) = \frac{0.104}{0.104 + 0.173}$$

$$= 0.375$$

$$\mathrm{P}(\omega_2 | \mathbf{x}) = \frac{0.173}{0.104 + 0.173}$$

$$= 0.625$$

このとき、患者はクラスω_2（無再発クラス）の一員と識別される。

図 2.4 に、マーカーx_1, x_2における分割$x_{1(1)}$，$x_{2(3)}$に対する条件付き確率を示す。図 2.4 では、マーカー1 と 2 でクラスω_1とクラスω_2の差異が明確に表わされている。

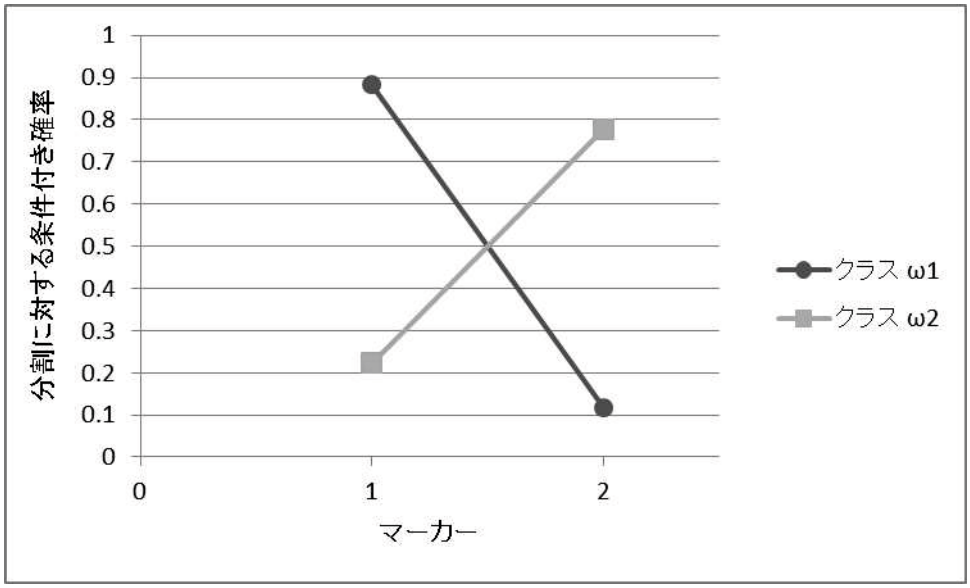


図 2.4 マーカーx_1, x_2における分割に対する条件付き確率

この図において、クラスω_1では$\mathrm{P}(x_{1(1)}|\omega_1)$の値が高く、$\mathrm{P}(x_{2(3)}|\omega_1)$の値は低い。逆にクラス$\omega_2$では$\mathrm{P}(x_{1(1)}|\omega_2)$の値が低く、$\mathrm{P}(x_{2(3)}|\omega_2)$の値は高くなっている。本論文では、各分割に入るサンプル数の割合を確率とみなし、Bayes 識別則の考えを用いて事後確率を計算している。2 クラスの事後確率の差は、図 2.4 に示すクラス間のマーカーの分割に入る患者の割合の差異を表わす「ある種の統計量」として考えられる。

次に、共に事後確率を基に識別を行う、離散 Bayes 識別器と従来の Bayes 識別器を比較する。従来の Bayes 識別器では、前述したようにサンプルは量的データを成分とする多次元の数ベクトルで表現され、その分布の統計情報は正規分布を仮定すると平均ベクトルと共分散行列にある。一般にサンプル数は少なく、次元数は高い。このとき共分散行列の推定誤差が大きく識別性能に影響を与え、極端な場合、共分散行列の逆行列が存在しないこともある。逆行列が存在しない場合は Bayes 識別器の設計が不可能となる。次に計算コストの観点か

らは、量的データのみを扱う Bayes 識別器では計算がベクトル間、ベクトルと行列間、あるいは行列間であり、次元数の増加に伴い計算コストは急激に増加する。一方、離散 Bayes 識別器では、全てが質的データ（記号）のみを扱うスカラー計算である。たとえ高次元であっても容易に計算ができる離散 Bayes 識別器の優位性は、実用性を考える上で極めて重要であると考える。

さて、これまで議論を簡単にするために、２クラス問題を対象としてきたが、本識別則は容易に多クラス問題へ拡張することができる。クラス数を$m(m \geqq 3)$として、クラスω_1、ω_2、…、ω_mが互いに排反で、かつその和集合が全集合とする。これは、パターンがm個のクラスのうちのいずれか一つのクラスに属することを意味する。このとき、マーカー$x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)}$が用いられると、クラス毎に条件付き確率$\mathrm{P}(x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)}|\omega_i)$、$i = 1,2, \ldots, m$ が表から計算され、事後確率$\mathrm{P}(\omega_i|\, x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)})$は

$$\mathrm{P}(\omega_i|x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)}) = \frac{\prod_{j=1}^{d} \mathrm{P}(x_{j(r_j)}|\omega_i)}{\sum_{k=1}^{m} \prod_{j=1}^{d} \mathrm{P}(x_{j(r_j)}|\omega_k)}$$

で与えられる。パターン$\mathbf{x} = [x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)}]$ は、その事後確率が最大のクラスを求め、

$$\mathrm{P}(\omega_k|x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{k(r_k)} = \max_i \mathrm{P}(\omega_i|x_{1(r_1)}, x_{2(r_2)}, \ldots, x_{d(r_d)})$$

であるならば、クラスω_kへ識別される。

2.4 識別手法及び評価指標

識別則の評価指標、また識別則の比較に用いる ROC 解析を説明する。

表 2.4 に癌の診断を例にした混合行列を示す。

A：本当は癌である症例を癌と診断した数
B：本当は非癌である症例を癌と診断した数
C：本当は癌である症例を非癌と診断した数
D：本当は非癌である症例を非癌と診断した数

表 2.4　混同行列

		正解	
		癌	非癌
診断結果	癌	A	B
	非癌	C	D

本論文で用いる評価指標を定義すると、以下となる[2, 3]。

①識別率 $= (A+D)/(A+B+C+D)$

識別率は全症例の中で正解ラベルが癌である症例を癌と、非癌である症例を非癌と正しく識別した割合で、この値が大きい程、性能が良いといえる。

②感度(sensitivity, recall) $= A/(A + C)$

感度は実際に癌である症例の中で正しく癌と識別した割合で、この値が大きい程、性能が良いといえる。

③特異度(specificity) $= D/(B + D)$

特異度は実際に非癌である症例の中で正しく非癌と識別した割合で、この値が大きい程、性能が良いといえる。

④precision = $A/(A+B)$

precision は癌と診断された症例の中で正しく癌であると識別した割合で、この値が大きい程、性能が良いといえる。

⑤F1 measure $= \frac{2\times recall\times precision}{recall+precision} = \frac{2A}{2A+B+C}$

F1 measure は感度と precision の調和平均で、この値が大きい程、性能が良いといえる。

⑥DOR（Diagnostic odds ratio）$= \frac{sensitivity\times specificity}{(1-sensitivity)\times(1-specificity)} = \frac{A\times D}{B\times C}$

DOR は診断オッズ比のことで、陽性尤度比（真陽性確率を偽陽性確率で割った値）を陰性尤度比（偽陰性確率を真陰性確率で割った値）で割った値で、この値が大きい程、性能が良いといえる。

⑦ROC 解析[4]：識別規則の識別性能を図示する際によく用いられる。ここで図 2.5 に例として ROC 図を示す。

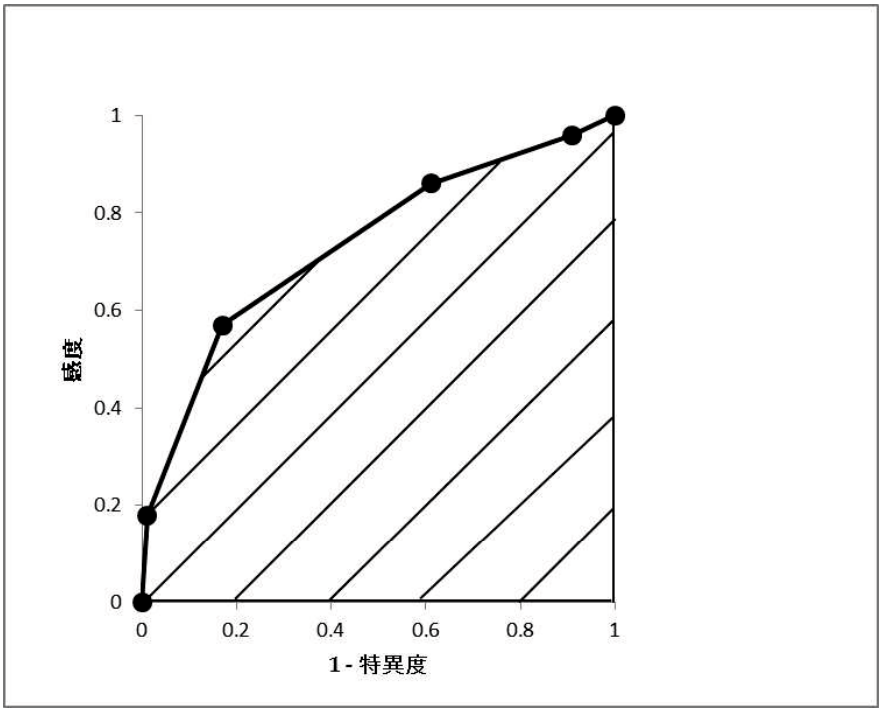


図 2.5 ROC 図

スコア式の値は、カットオフ値を設けると、カットオフ値との大小関係により診断を行い、感度と特異度の組を求めることができる。さらにカットオフ値を変えて感度と特異度の組を複数求めることもできる。その後、診断の結果を縦軸を感度、横軸を 1－特異度として、感度と 1－特異度の組をプロットし、プロット点を結ぶことで ROC 曲線を作成する。ROC 曲線が左上に近ければ近い程、プロット点における識別性能は高いといえる。また、ROC 曲線を用いて複数の識別則の識別性能を比較する場合は、AUC (Area Under the Curve) を用いる。AUC は図 2.4 中の斜線部で示した部分で曲線の下部である。この AUC の値が大きい程、性能は良いといえる。

参考文献

1) 浜本義彦, 統計的パターン認識入門, 森北出版, 2009.

2) D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, Journal of Machine Learning Technologies, Vol.2, No.1, pp.37-63, 2011.

3) A.S. Glas et al., The diagnostic odds ratio: a single indicator of test performance, Journal of Clinical Epidemiology, Vol. 56, No. 11, pp.1129-35, 2003.

4) R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, Second Edition, John Wiley & Sons, 2001.

第 3 章　離散 Bayes 識別則による個別化医療への展開

3.1 はじめに

本論文では、第２章で提案した離散 Bayes 識別器を肝癌の早期再発の予測、早期胃癌におけるリンパ節転移の予測、大腸癌における抗癌剤と免疫療法の併用効果の予測、漢方薬の処方の問題について適用し、その有用性について検討する。以下、４つの個別化医療問題に対する適用結果を示す。

3.2 肝癌の早期再発の予測

肝癌は、難治性の癌の一つとして、その克服が国民的課題となっている。その難治性は、手術で癌を完全に切除しているのにもかかわらず、肝癌再発の高率にある[1]。この再発を精度良く予測できれば効率的な先制医療が実施でき、無再発なら抗癌剤の投与が不要となり、また CT 検査の必要もなくなる。その結果、医療費の抑制にもなる。

前述したように、癌治療の困難さは、たとえ癌種が同じであっても患者個々によって癌が異なる多様性にあり、血液検査や CT 検査などの多種多様なマーカーも癌の一側面しか診ていなく、現時点では決め手となるマーカーがない。そのため、これまでマーカーを基に肝臓の状態を表わすスコア式として Tokyo Score, Modified JIS, TNM 分類が提案されているが、いずれも臨床現場の要求に応えきれていない。その原因は、上述のようにスコア式に用いるマーカーの組は医師による試行錯誤の末に得られたものである点にある。一方、機械学習による癌の診断では、統計的パターン認識の Bayes 識別器と同じく、質的データには適用できないという問題があった。この問題に対処するために、質的デ

ータにも適用できる離散 Bayes 識別器を提案する。量的データについては、しきい値処理により量的データを質的データに変換し、全てのデータを質的データに統一する。そして、それらを成分とするパターンベクトルとして患者を表わす。次に再発予測の問題を、患者が再発するクラスのパターンか、あるいは再発しないクラスのパターンかを識別する２クラス問題として定義する。この２クラス問題に対し、第２章の離散 Bayes 識別器により、肝癌再発の有無を高精度で識別する。また、マーカー数の多さ、つまりパターンベクトルの高次元の問題に対処するため、仮想サンプルを用いたリサンプリング技術[2]に基づく特徴選択により最適マーカーの組を選択する。

3.2.1　マーカー探索

どのマーカーを識別器に用いるかは、識別において本質的である。これは、統計的パターン認識における特徴選択の問題[3]である。ここでは、M 個の候補マーカーの中から識別に有用な d 個のマーカーの組合せを選択する特徴選択問題を解く方法を説明する。少数の訓練サンプル下におけるマーカー選択は、選択されたマーカーを用いる識別器が訓練サンプルの完全な識別を可能にする一方で、新しいパターンではうまく動作しないことが多い。これをオーバーフィッティング問題[4]と呼ぶ。オーバーフィッティング問題を回避するために、利用可能な少数の訓練サンプルからリサンプリング技術により生成された仮想サンプルを用いることによって、最適なマーカーの組合せを求めることができる。このアイデアは、ブートストラップ技術[5]に由来する。特徴選択の流れを図 3.1 に示す。ここで感度と特異度を定義しておく。感度とは、再発患者の中で正しく再発と識別された患者の割合である。一方、特異度とは、無再発患者の中で

正しく無再発と識別された割合である。

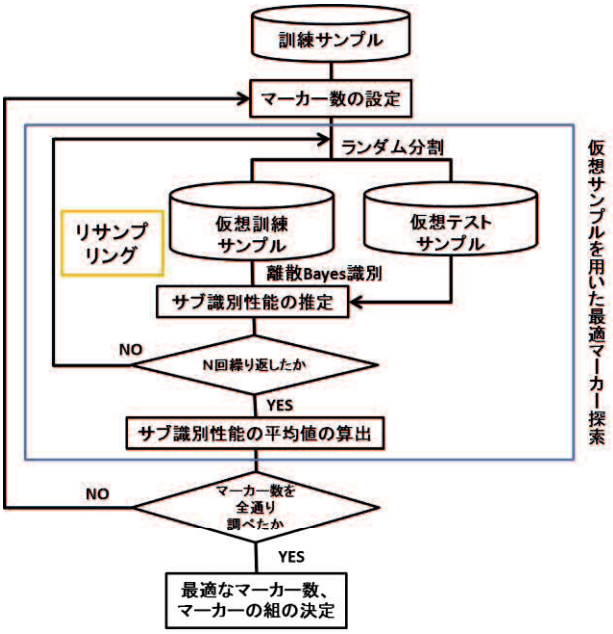


図 3.1　最適マーカーの選択

初めに、訓練サンプルを、仮想訓練サンプルと仮想テストサンプルへランダムに 2 分割する。そして探索開始のマーカー数を決定し、評価対象のマーカーの組を一つ選択する。次に、離散 Bayes 識別則に基づき、仮想訓練サンプル上で離散 Bayes 識別器を設計し、仮想テストサンプルに対する条件付き確率 $P(\mathbf{x}|\omega_i)$を計算する。その後、事後確率によって仮想テストサンプルの識別を行う。この識別では、まず仮想テストサンプルの測定値がいま評価対象となっているマーカーのどの分割に属するかを調べる。次に属する分割に対応する条件付き確率を用いて事後確率を求め、最大事後確率のクラスへ仮想テストサンプルを識別する。以上の処理を独立に N 回繰り返す。仮想テストサンプルに対するサブ識別性能としての感度及び特異度の平均は、N 回の試行から推定される。次に、同じマーカー数の全ての組合せを評価する。ここで、本論文では再発予測を重視しているため、感度が高いことが要求される。そこで、当該マーカー数において仮想テストサンプルに対する平均特異度 0.5 以上の制約条件のもと

で平均感度を最大にする最適なマーカーの組を最適なマーカーの組合せの候補として選択する。次にマーカー数を一つ増加させ同様な処理をマーカー数 M－1 まで繰り返す。そして最適なマーカーの組合せの候補の中で、平均感度が最大となり、よりマーカー数の少ない組を最終的に、最適なマーカー数 d とその組として決定する。

3.2.2 実験方法

まず、用いるデータを説明する。症例は山口大学医学部附属病院から提供され，手術ですべての肝癌を取り除いた患者から得られたものである．この中に肝癌が 1 年以内に再発した患者が５７名，無再発の患者が１７７名含まれている．肝癌は感染するウイルスのタイプによって C 型肝癌、B 型肝癌等に分類される。用いた肝癌のウイルスタイプを表 3.1 に示す。なお、提案手法は、前述の Tokyo Score, Modified JIS や TNM 分類と同様に、ウイルスのタイプに依存しない方法である。

本論文では ALB, 腫瘍数×腫瘍サイズ[6], ICG, vp, vv、血小板, PT, ビリルビン、分化度、Liver damage、計 10 マーカーをマーカー候補とし、この中から探索して、最適マーカーの組合せを決定し、その組合せを離散 Bayes 識別器に用いる．この表 3.2 について説明する。各マーカーに対するカットオフ値は予め医学的に決められている。例えば ALB というマーカーでは 3.5 よりも大であるか否かで再発患者と無再発患者それぞれを分割する。再発患者について言えば、3.5 よりも大が 15 名、3.5 以下が 14 名の計 29 名となっている。どのマーカーにおいてもマーカー内では分割された再発患者の総数は 29 名である。

表 3.1 (a) 訓練サンプルの内訳

ウイルスタイプ	1年再発	1年無再発
B	6	16
C	18	56
(−): *B型またはC型でないもの*	5	17
Total number of samples	29	89

表 3.1 (b) テストサンプルの内訳

ウイルスタイプ	1年再発	1年無再発
B	6	16
C	17	55
(−): *B型またはC型でないもの*	5	17
Total number of samples	28	88

表 3.2 各クラスにおけるマーカーの分割の内訳

	サンプル数	29	89
マーカーの分割		**一年以内再発**	**一年無再発**
$x_{1(1)}$	ALB > 3.5	15	60
$x_{1(2)}$	ALB ≤ 3.5	14	29
$x_{2(1)}$	**腫瘍数 × 腫瘍サイズ** < 4	6	47
$x_{2(2)}$	**腫瘍数 × 腫瘍サイズ** 4~9	11	29
$x_{2(3)}$	**腫瘍数 × 腫瘍サイズ** > 9	12	13
$x_{3(1)}$	vp +	10	18
$x_{3(2)}$	vp -	19	71
$x_{4(1)}$	ICG < 15	14	44
$x_{4(2)}$	ICG ≥ 15	15	45
$x_{5(1)}$	vv +	9	16
$x_{5(2)}$	vv -	20	73
$x_{6(1)}$	**血小板数** ≥ 10	16	70
$x_{6(2)}$	**血小板数**< 10	13	19
$x_{7(1)}$	PT ≥ 80	18	67
$x_{7(2)}$	PT < 80	11	22
$x_{8(1)}$	Bilirubin < 1	17	61
$x_{8(2)}$	Bilirubin ≥ 1	12	28
$x_{9(1)}$	**分化度** non por	25	79
$x_{9(2)}$	**分化度** por	4	10
$x_{10(1)}$	Liver damage A	15	60
$x_{10(2)}$	Liver damage B	14	29

次に識別性能の評価方法について説明する。識別性能の推定は、訓練サンプルとテストサンプルの独立性を保つ Hold-out 法[7]を用いる。再発５７サンプル，無再発１７７サンプルをランダムに半分ずつに分割し，その一方を訓練サンプルとし、残りをテストサンプルとする。具体的には再発の２９訓練サンプル数、無再発の８９訓練サンプル数、一方、再発の２８テストサンプル、無再発の８８テストサンプルとした。識別器の学習と評価の流れを図 3.2 に示す。識別器の設計および評価のフローが示されている最初にマーカーの最適な組合せを決定する。識別器の識別性能の評価値としてテストサンプルに対する識別率、感度、特異度、Youden_index（感度＋特異度－1.0）[8]、F1 measure、診断オッズ比[9,10]を採用する．これらの値は高いほど識別器の識別性能が高いといえる．ここで感度とは，再発患者の中で正しく再発と識別された患者の割合を意味する．一方，特異度とは，無再発の患者の中で正しく無再発と識別された割合とする．本論文では、早期再発の予測を目的とするために、特異度が 0.5 以上の制約条件の下で感度を評価する．

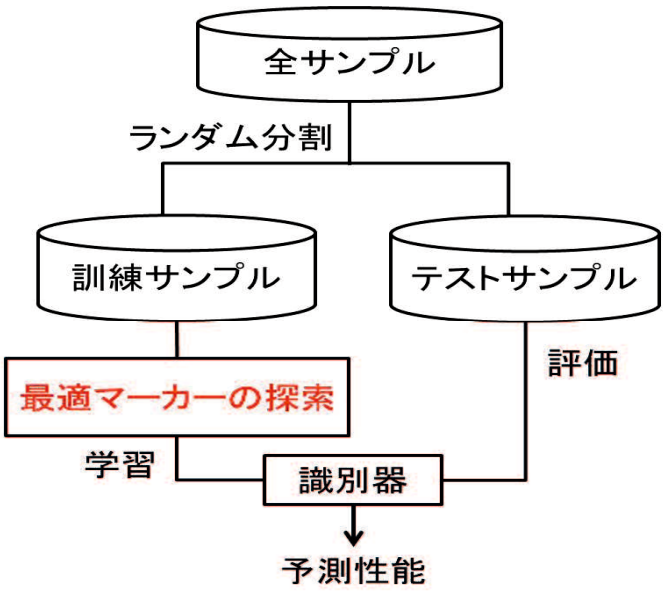


図 3.2　識別器の学習と評価の流れ

一般に訓練サンプル数が増加すると、識別器の識別性能は向上する。そのため、どのくらいの訓練サンプル数が識別器の設計に必要かは識別器の設計者にとって興味のあることである。そこで、訓練サンプル数を増加させた部分集合として、図 3.3 に示す 6 訓練サンプルの部分集合の系列が、以下の関係を有すると仮定する。

$S_1 \subset S_2 \subset S_3 \subset S_4 \subset S_5 \subset S_6$

このとき、部分集合は入れ子構造をとる。最初の部分集合 S_1 は再発の 5 訓練サンプルと、無再発の１５訓練サンプルとして、計２０個の訓練サンプルを要素とする。次の部分集合 S_2 は真に再発の訓練サンプルを 1 個、無再発の訓練サンプルを 2 個追加し、２３個の訓練サンプルを要素とする。このように真に訓練プンプルを増加させた部分集合それぞれで離散 Bayes 識別器を設計し、得られる 6 個の離散 Bayes 識別器によりそれぞれ同一のテストサンプルに対する識別性能を求める。以上の試行を独立に３０回繰り返して、訓練サンプル数増加の識別性能への影響を調べる。

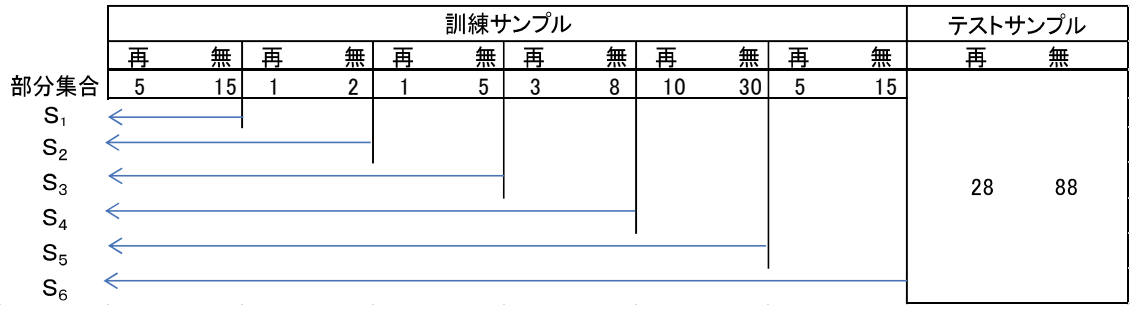


図 3.3　訓練サンプル部分集合間の関係

また、前述したように、離散 Bayes 識別器は、識別にスカラー計算だけ行うので、計算上のメリットがある。これを明らかにするために、116 個の実テストサンプルをコピーして 1160000 個の人工テストサンプルを用意する。そして表 3.3 に示すマーカーの組合せを用い、マーカー数を 3 から 6 まで一つずつ変え、各マーカー数における最適マーカーの組の候補を用いて人工テストサンプルに対する識別時間を調べる。識別時間は、識別処理の開始から終了までの時間とし、clock 関数を用いて計測する。

次に、既存のスコア式との性能比較を行う。前述した通り、識別器の識別性能の評価値としてテストサンプルに対する識別率、感度、特異度、Youden_index (感度＋特異度-1.0) [8]、F1 measure、診断オッズ比[9,10]を採用する。また、性能比較のために ROC 解析[4]も行う。

比較対象のスコア式について説明する。Tokyo Score, Modified JIS, TNM 分類は、用いる各マーカーの測定値に対して、医師が定めたカットオフ値によりスコア値が与えられる。次に、スコア値の総和をとって、総スコア値とする。この総スコア値に対してカットオフ値により患者は診断される。予測方法は以下の図 3.4 を用いて説明される。予測の規則として、総スコア値が、cut-off 値未満ならば無再発，cut-off 値以上ならば再発と予測する。この例では、総スコア値が 0～8 をとり、cut-off 値が 1 と設定しているため、総スコア値が 0 のときのみ患者を無再発と予測し、1 以上のときは患者を再発と予測する。

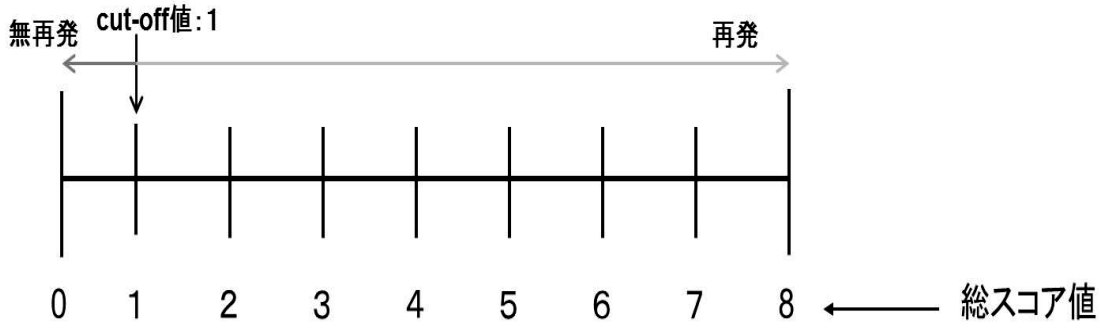


図 3.4 スコア式を用いた予測方法

例えば Tokyo Score ではアルブミン、ビリルビン、腫瘍サイズ、腫瘍数の 4 つのマーカーが用いられている。ある患者のアルブミン値が 3.0(g/dl)、ビリルビン値が 1.5(g/dl)、腫瘍サイズが 1.0(cm)、腫瘍数が 4 であった場合、各マーカーのスコア値はそれぞれ 1 点, 1 点, 0 点, 2 点となり、総スコア値は 4 点となる。仮に cut-off 値を 2 とすれば、その患者は再発すると診断される。

3.2.3 実験結果

表 3.3 は、マーカーの数当たり 100 回のリサンプリングから得られたマーカーの候補の組合せおよびそれらの識別性能を示す。表中のダッシュ記号は、特異度の制約条件を満たす組合せが存在しないことを示す。特異度 0.5 の制約条件下での感度に基づいて、腫瘍数×腫瘍サイズ、vp、ICG および Liver damage の 4 つのマーカーの組合せを最適として同定した。

表 3.3　訓練サンプルに対するマーカー数毎の最適なマーカーの組合せの候補とその識別性能

マーカー数	感度	特異度	Youden Index	マーカーの組								
3	0.79	0.50	0.29	腫瘍数＊腫瘍サイズ	vp	Liver damage						
4	0.80	0.50	0.30	腫瘍数＊腫瘍サイズ	vp	ICG	Liver damage					
5	0.75	0.50	0.25	ALB	腫瘍数＊腫瘍サイズ	vv	分化度	Liver damage				
6	0.74	0.51	0.24	ALB	腫瘍数＊腫瘍サイズ	vp	ICG	分化度	Liver damage			
7												
8												
9												

次に訓練サンプル数と感度との関係を図 3.5 に示す。エラーバーは感度の 95%信頼区間を示す。この信頼区間の幅は訓練サンプル数が少ないときほど広く、サンプルが増えるごとに縮小していることが確認された。ここで、信頼区間の幅は信頼性を表わし、幅が狭いほど信頼性が高いと考えられる。また、訓練サンプル数が 80～100 において信頼区間がオーバーラップしている。このことから識別性能は飽和しているといえる。

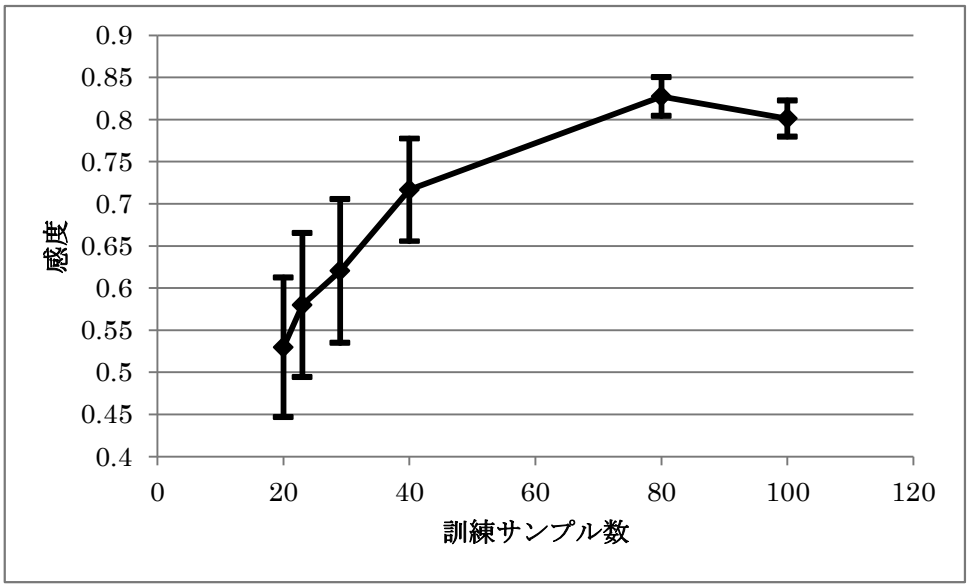


図 3.5　訓練サンプル数と感度との関係

次に離散 Bayes 識別器と既存の肝臓スコア式の性能を比較した。識別器は、F1 measure や診断オッズ比[9,10]などのよく知られた指標、そして ROC 解析によって評価した。それらの結果をそれぞれ表 3.4 と図 3.6 に示す。

表 3.4 提案手法と既存のスコア式との性能比較

	提案手法	Modified JIS with 3	TNM 分類 with 2	Tokyo score with 2
識別率	0.58	0.77	0.34	0.49
感度	0.86	0.57	0.96	0.71
特異度	0.49	0.83	0.14	0.42
F1 measure	0.49	0.54	0.41	0.40
Youden index	0.35	0.40	0.10	0.13
Diagnostic odds ratio	5.73	6.49	4.26	1.81

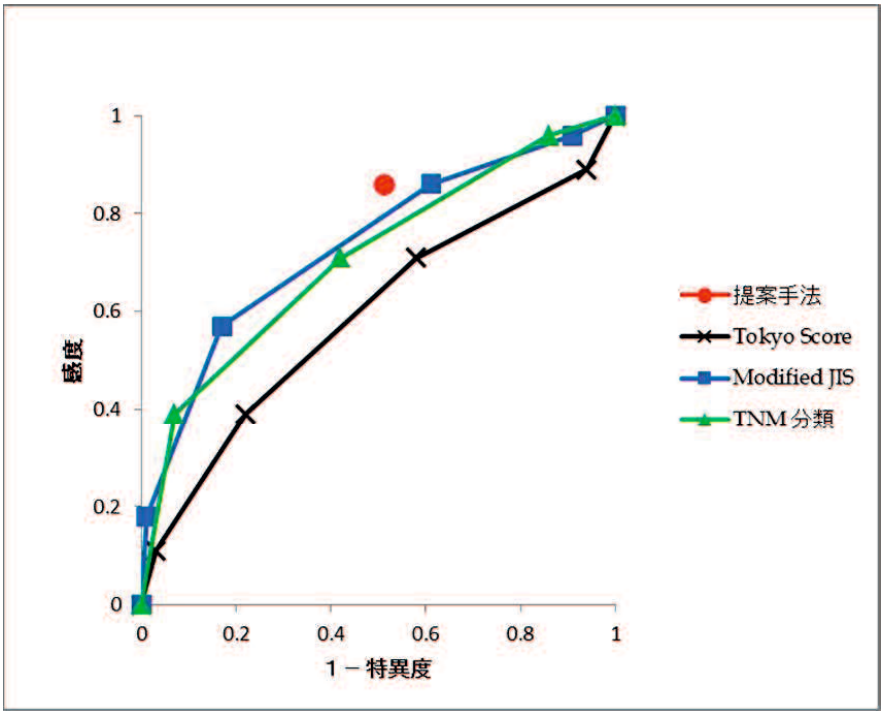


図 3.6　ROC 曲線

最後にマーカー数を 3 から 6 に 1 つずつ変えて求めたマーカーの数と識別時間（CPU time）との関係を図 3.7 に示す。図 3.7 より、識別時間は指数関数のように急激に増加するのではなく、線形関数のように増加していることが確認された。

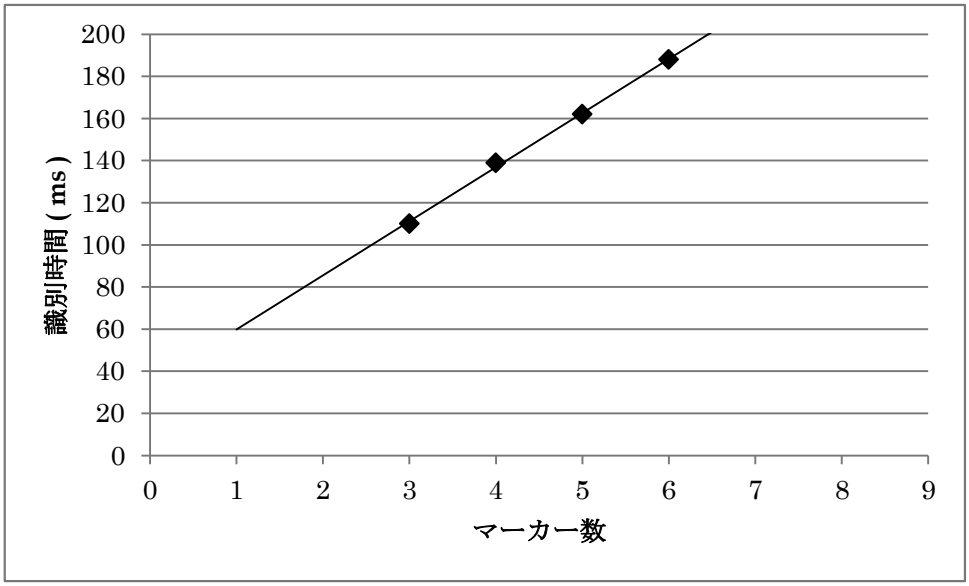


図 3.7 マーカー数と識別時間との関係.

3.2.4 考察

実験により、１０個のマーカーの候補の中から独自の特徴選択により選択された４マーカー（腫瘍数×腫瘍サイズ、vp、ICG、Liver damage）の組合せを同定した。次に、訓練サンプル数の識別性能への影響も調べ、図 3.5 に示すように訓練サンプル数を増加させるにつれて識別性能は収束していく様子が見て取れ、これより訓練サンプル数は満たされていることが確認できた。次に、識別に要する計算コストについて調べた。マーカー数を増加させると図 3.7 に示すように、識別時間は線形のオーダーで増加した。よって、メチル化や遺伝子のように数十万、数万のマーカー候補があるビックデータに対して、本手法は計算コストの点でメリットがあることを示した。また、同定したマーカーの組を用いたテストサンプルに対する離散 Bayes 識別則の識別性能は、感度 0.86、特異度 0.49 であった。この識別性能を既存の代表的な３つの肝臓スコア式と比較した。識別において癌の再発を見逃さないという理由から、特異度を一定のレベルにキープして感度が高いことを重視している。識別性能の比較結果より、離散 Bayes 識別則は、既存のスコア式ほど特異度を低下させることなく、再発の早期予測に重要な指標である高い感度を達成できたといえる。

識別性能である識別率, F1 measure では、再発テストサンプル数と無再発テストサンプル数が同じであることを前提としているが、本論文のテストサンプル数はこれに反している。そこで、再標本化法を用いてサンプル数を仮想的に一致させ、人工の 28 個の無再発テストサンプルと実の 28 個の再発テストサンプルを用いて再評価した。その結果を表 3.5 に示す。上段は平均値、下段は 95% 信頼区間を示す。仮に２クラスのテストサンプル数が同数であれば、識別率, F1 measure においても本手法は既存のスコア式に見劣りしない。次に図 3.6 の ROC 解析では、既存のスコア式の ROC 曲線に比べ、本手法はそれらの上に位

置した。これは既存のスコア式よりも優れていることを示唆している。

この他に離散 Bayes 識別器の長所として、既存のスコア式が特定のマーカーを必要としているのに対し、提案手法は患者毎に与えられたマーカー候補の中から最適なマーカーの組合せを求めることができる点がある。このため、既存のスコア式は指定されたマーカーのデータに欠損がある場合には利用できないが、提案手法では患者が有する検査データの中から最適なマーカーの組合せを選択して用いることができる。更にどのマーカーを追加すれば識別性能が向上するかを医師に示すこともできる。このように離散 Bayes 識別則によれば、最良の個別化医療が期待できる。

表 3.5 リサンプリング法による
テストサンプル（再発 28 サンプルと無再発 28 サンプル）の識別結果

	提案手法	Modified JIS with 3	TNM 分類 with 2	Tokyo score with 2
識別率	0.67 [0.66, 0.69]	0.70 [0.69, 0.71]	0.55 [0.54, 0.56]	0.57 [0.55, 0.58]
感度	0.86	0.57	0.96	0.71
特異度	0.49 [0.46, 0.52]	0.83 [0.81, 0.85]	0.14 [0.12, 0.16]	0.42 [0.39, 0.44]
F1 measure	0.73 [0.72, 0.73]	0.66 [0.65, 0.67]	0.68 [0.68, 0.69]	0.62 [0.62, 0.63]
Youden index	0.35 [0.32, 0.37]	0.40 [0.38, 0.42]	0.10 [0.08, 0.12]	0.13 [0.11, 0.16]
Diagnostic odds ratio	6.03 [5.39, 6.67]	7.88 [6.22, 9.53]	4.62 [3.86, 5.38]	1.95 [1,74, 2.15]

3.3 早期胃癌におけるリンパ節転移の予測

胃癌は多くの国民がかかる癌の一つである。2012 年の男女合計での罹患数は部位別で 2 位、2014 年の死亡数は部位別で 3 位となっている[12]。しかし、胃癌は早期発見ができれば治る癌となった。早期発見には胃カメラの検査があり、早期胃癌であれば内視鏡的粘膜層剥離術(以下、ESD 治療) が有効となっている。ESD 治療は従来手法の EMR（内視鏡的粘膜切開術）の問題点であった切除できるサイズが小さかった点と切除が分割して行われるが故に、癌が残ってしまう点を改善した手法である。大きく癌を切除することが可能となったためにこれらの問題点を大幅に改善できたといえるが、それでも癌の深さなどの原因からリンパ節に転移する例もある。もし早期胃癌に対する ESD 治療後にリンパ節転移があれば更に外科手術が必要である。しかし、現状は転移の特異度が低い。つまりリンパ節転移がないにも関わらず、必要のない手術が行われている。そのためリンパ節転移の予測は、患者にとって重要である。本論文では、ESD 治療後に利用可能なデータのみを用いて、リンパ節転移の有無を高精度で予測することを目的とする。

3.3.1 実験方法

最初に用いるデータを説明する。症例は山口大学医学部附属病院から提供され，ESD 治療後 3 年以上の経過観察期間を有するか、あるいは追加外科切除が行われた 423 名の患者から得られたデータである。この中にリンパ節転移のある患者が 9 名，リンパ節転移のない患者が 414 名含まれている。

マーカー候補として、医学的観点から医師によって定められた性別、肉眼型、深達度、分化度、ｌｙ（リンパ管侵襲)、ｖ（静脈侵襲）UI（潰瘍)、腫瘍径の計８マーカーを用いる．各マーカーの分割と患者数を表 3.6 に示す。

表 3.6　各マーカーの分割と患者数

マーカーの分割	リンパ節転移あり	リンパ節転移なし
$x_{1(1)}$ 男性	7	336
$x_{1(2)}$ 女性	2	78
$x_{2(1)}$ 隆起型	7	157
$x_{2(2)}$ 陥凹型	2	257
$x_{3(1)}$ M, SM1	1	375
$x_{3(2)}$ SM2	8	39
$x_{4(1)}$ 分化型	5	374
$x_{4(2)}$ 混在・未分化型	4	40
$x_{5(1)}$ リンパ管侵襲なし	1	381
$x_{5(2)}$ リンパ管侵襲あり	8	33
$x_{6(1)}$ 静脈侵襲なし	4	399
$x_{6(2)}$ 静脈侵襲あり	5	15
$x_{7(1)}$ 潰瘍なし	8	373
$x_{7(2)}$ 潰瘍あり	1	41
$x_{8(1)}$ 腫瘍径30mm未満	7	348
$x_{8(2)}$ 腫瘍径30mm以上	2	66

次に最適マーカー探索法と識別性能の評価方法を説明する。最適マーカーの組合せを探索するためと、得られた最適マーカーの組を評価するために leave-one-out 法[13]を 2 回用いる。leave-one-out 法は、少ないサンプルしか利用できない状況下で、識別器の設計を行い、精度の高い識別性能を推定するために考案された手法である。leave-one-out 法では n 個のサンプルが利用できる場合、その中から一つテストサンプルとして選出し、残りの$n-1$個のサンプルを訓練サンプルとする。そして訓練サンプルを用いて離散 Bayes 識別器を設計し、その離散 Bayes 識別器でテストサンプルを識別する。以上の処理を各サンプルがただ一度限りテストサンプルとして用いられるまで繰り返す。実験の流れを図 3.8 に示す。まず最適マーカーの組合せの探索のために leave-one-out 法に従って 423 症例の中から 1 例をテストサンプルとして選出し、残りの 422 例を訓練サンプルとする。

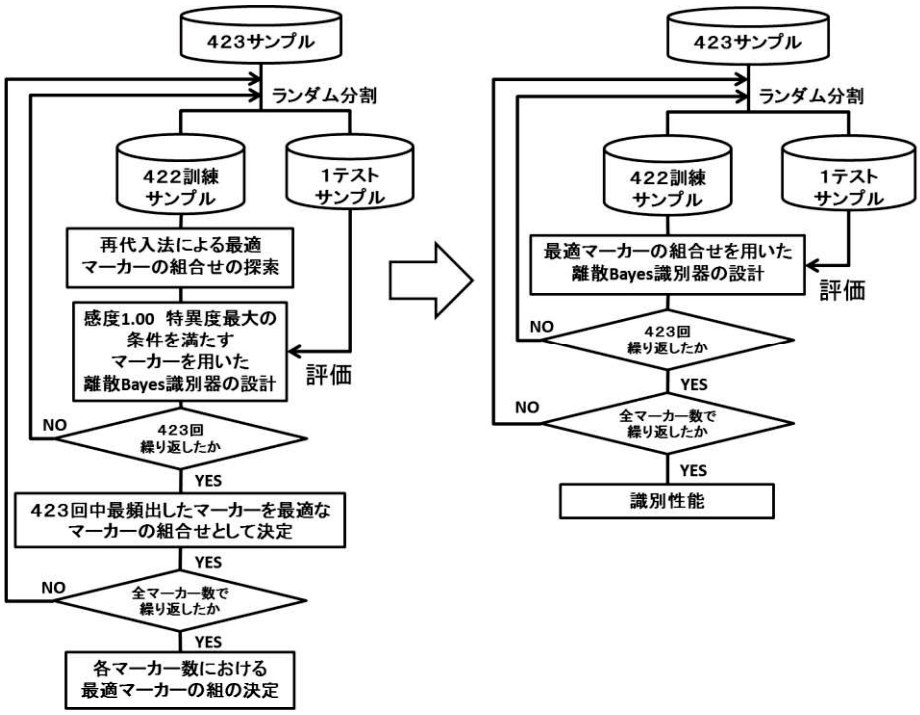


図 3.8 実験の流れ

422 例を訓練サンプルとした後、leave-one-out 法の代わりに再代入法[14]を用いてマーカーの組合せを評価する。再代入法とは、利用できるサンプル全てを訓練サンプル、またテストサンプルとして用いる手法である。再代入法により、422 のサブ訓練サンプルで離散 Bayes 識別器を設計し、422 のサブテストサンプルに対する感度と特異度を求める。この処理を全てのマーカーの組合せが評価されるまで繰り返す。具体的にはマーカー数を 3 から 8 まで変え、全ての組合せ、 ${}_8C_3+{}_8C_4+{}_8C_5+{}_8C_6+{}_8C_7+{}_8C_8=219$ 通り、を評価する。この 219 通りの中からサブテストサンプルに対する感度が 100%かつ特異度が最大となるマーカーの組合せをマーカー毎に一つ選択する。 選択されたマーカーの組合せで上記 422 個の訓練サンプルを用いて離散 Bayes 識別器を設計し、テストサンプルを識別する。以上の処理を各サンプルがただ一度限りテストサンプルとして用いられるまで, すなわち 423 回独立に繰り返す。その結果、423 回マーカーの組合せが選択される。ここまでの流れが図 3.7 中の左部のフローチャートに対応しており、以降の流れが右部に対応する。423 マーカーの組の中で最頻出するマーカーの組は、リンパ節転移を予測するために必須なマーカーが含まれていると考えられる。そこで最頻出するマーカーの組合せを固定して leave-one-out 法に従って離散 Bayes 識別器を設計し、 テストサンプルに対する感度, 特異度, 識別率を評価する。ここで感度とは、リンパ節転移のあった患者を正しく転移ありと予測できた割合を意味する。一方、特異度とは、リンパ節転移がなかった患者を正しく転移なしと予測できた割合とする。

3.3.2 実験結果と考察

表 3.7 にマーカー毎に最頻出したマーカーの組合せとその組を固定した場合のテストサンプルに対する識別性能を示す。

表 3.7　制約条件「訓練サンプル上で感度 1.00」のもと、
特異度が最大となるマーカーの組合せの中で、最頻出したマーカーの組合せ

マーカー数	マーカーの組								識別性能	感度	特異度	leave-one-outの検証行423回の中で選択された回数
2	ly								0.90	0.89	0.90	1
3	類似構造	v							0.86	1.00	0.86	415
4	類似構造	類似構造	v						0.86	0.89	0.86	418
5	側性	類似構造	類似構造	ly	v				0.91	0.89	0.91	421
6	側性	類似構造	類似構造	ly	v	UI			0.89	0.89	0.89	360
7	側性	類似構造	類似構造	分化型	ly	UI	v		0.90	1.00	0.89	418
8	側性	類似構造	類似構造	分化型	ly	v	UI	臨終期癌	0.88	1.00	0.87	423

感度が 1.00 の制約条件の下で、特異度が最大となり、またマーカー数はできるだけ少ないという観点から、マーカー数 3 の深達度、ly、v の組が識別に最適であると考えた。なお、マーカー数が 2 のとき、423 回の独立試行の中で 419 回制約条件が満たされなかった。

この深達度、ly、v を最適なマーカーの組合せとして用いた場合のテストサンプルに対する予測結果を表 3.8 に示す。深達度、ly、v の全ての組合せ総数は 8 であり、それぞれのケースについてのクラスω_1、ω_2の事後確率を示している。例えばケース 1 の場合、$P(\omega_1|x)$の平均値は 0.334、$P(\omega_2|x)$の平均値は 0.666 となっており、$P(\omega_2|x)$の方が事後確率は高い。離散 Bayes 識別器では事後確率の高い方のクラスへサンプルを識別するため、ケース 1 に該当する 354 サンプルは全てω_2と識別される。表 3.8 に示すように全 423 サンプル中、363 サンプルを正しく識別した。

表 3.8　テストサンプルに対する予測内容

ケース	X_1: 深達度	X_2: ly	X_3: v	$P(\omega_1 \mid x)$の平均値	$P(\omega_2 \mid x)$の平均値	判定	Bayes誤識別率平均値	感度	特異度
1	M、SM1	なし	なし	0.334	0.666	転移なし	0.334	–	1.00(354/354)
2	M、SM1	なし	あり	0.768	0.232	転移あり	0.232	–	0.00(0/1)
3	M、SM1	あり	なし	0.615	0.385	転移あり	0.385	1.00(1/1)	0.00(0/17)
4	M、SM1	あり	あり	0.866	0.134	転移あり	0.134	–	0.00(0/3)
5	SM2	なし	なし	0.580	0.420	転移あり	0.420	1.00(1/1)	0.00(0/20)
6	SM2	なし	あり	0.853	0.147	転移あり	0.147	–	0.00(0/6)
7	SM2	あり	なし	0.869	0.131	転移あり	0.131	1.00(2/2)	0.00(0/8)
8	SM2	あり	あり	0.540	0.460	転移あり	0.460	1.00(5/5)	0.00(0/5)
						誤り総数		0	60

3.4 大腸癌における抗癌剤と免疫療法の併用効果の予測

癌を治療する方法として、手術療法、化学療法、放射線療法があり、これらは三大療法と呼ばれている。この他に第四の療法として免疫療法[15]が近年注目されている。免疫療法とは体の免疫力を上げることで癌細胞などの異物を排除する治療法である。昨今の癌治療では抗癌剤、または放射線治療と免疫治療を併用することにより副作用の軽減などの相乗効果があると期待される。しかし、この免疫治療と用いる抗癌剤の効果を投与前に正確に予測するマーカーは見つけられておらず、個別化医療のために予測に用いるマーカーを発見することが望まれている。抗癌剤の効果を投与前に予測できれば、患者の身体的または金銭的負担を軽減することができる。すなわち個別化医療を実現できる。本論文ではマイクロ RNA を用いた離散 Bayes 識別則によりに抗癌剤と免疫治療との併用効果を予測する。

3.4.1 実験方法

最初に用いるデータを説明する。症例は山口大学医学部附属病院から提供され，免疫治療と抗癌剤治療の併用後の患者から得られたものである。経過観察として最大でも 6 年以内の患者が対象となっており、この中に免疫治療後 2 年以内に死亡した患者が 7 名、２年以上生存した患者が 8 名含まれている。最適なマーカーの候補となるマイクロ RNA は 811 マーカーである。この 811 マーカーの中から予測に有用な最適マーカーを選択する。

初めに、15 症例(2 年以内死亡の 7 症例、2 年以上生存の 8 症例)のうち、典型的な症例として生存期間の長い方から 4 症例を効果ありクラスの訓練サンプルとし、生存期間の短い方から 4 症例を効果なしクラスの訓練サンプルとして取り出す。残りの 7 症例はいずれも予測が困難な症例で、これらを全てテストサンプルとする。マーカー数 d を 2 から 5 へと一つずつ変え、各マーカー数において離散 Bayes 識別則に基づき、訓練サンプルを用いて各マーカーの組 $\mathrm{x} = (x_1, x_2, \cdots, x_d)$ に対する条件付き確率$\mathrm{P}(\mathrm{x}|\omega_i)$を計算し、事後確率$\mathrm{P}(\omega_i|\mathrm{x})$を求めて、事後確率が最大のクラスへ訓練サンプルを識別する。

マーカー探索であるが、訓練サンプルに対する識別率が最も大きいマーカーの組合せを最適とする。このとき識別率が同じであれば、訓練サンプル上で Δ（δ_iの総和）を特徴評価関数と定義し、これを最大にするマーカーの組合せを特徴選択して、最適マーカーの組を用いて離散 Bayes 識別器を設計した。ここでδ_iは

$$\delta_i = \begin{cases} |\mathrm{P}(\omega_1|\mathrm{x}) - \mathrm{P}(\omega_2|\mathrm{x})|, & if\ \mathrm{x}\text{が正識別} \\ -|\mathrm{P}(\omega_1|\mathrm{x}) - \mathrm{P}(\omega_2|\mathrm{x})|, & otherwise \end{cases}$$

であり、 Δ を

$$\Delta = \sum_{i=1}^{n} \delta_i$$

と定義し、n は訓練サンプル数を表す。ここで、図 3.9 にδと識別境界線までの垂直距離 L との関係を示す。離散 Bayes 識別則では事後確率が大きいクラスへパターンを識別する。従って縦軸と横軸の値が一致する、つまり原点を通る 45° の直線が識別境界線となる。

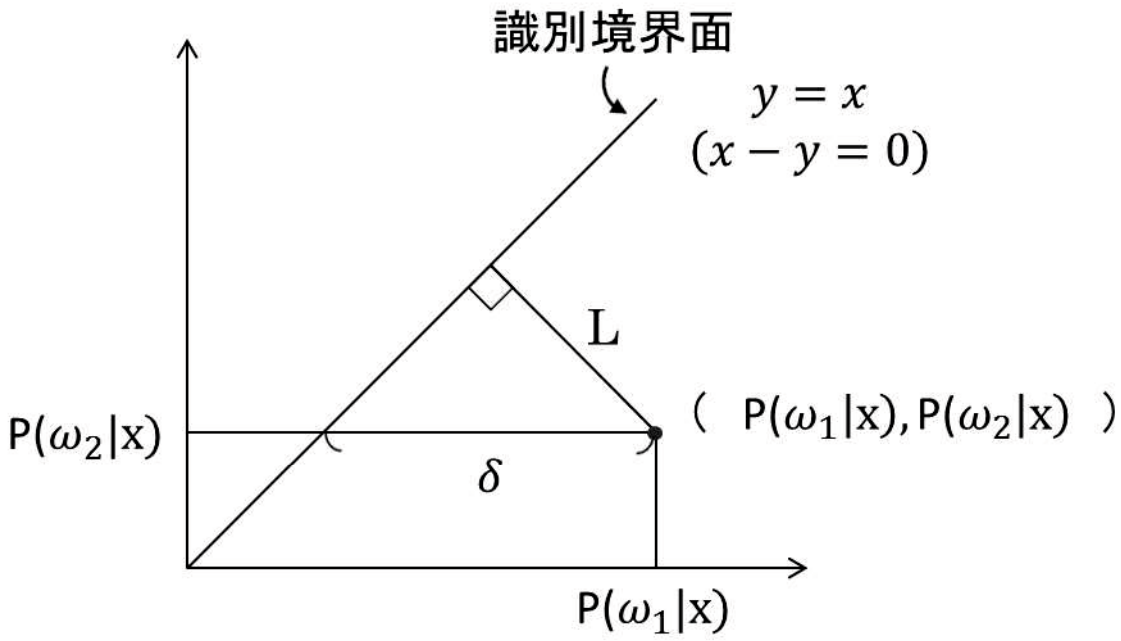


図 3.9　δ と L の関係

$$\delta = |P(\omega_1|x) - P(\omega_2|x)|$$

$$L = \frac{|P(\omega_1|x) - P(\omega_2|x)|}{\sqrt{1+1}}$$

$$= \delta/\sqrt{2}$$

δの値が大きいということは L の値が大きいことであり、L の値が大きいことはテストサンプルが識別境界線から遠く離れて確実に識別されていることを意味する。

各マーカーの組毎に設計された離散 Bayes 識別器を用いてテストサンプルを識別する。識別指標は、感度、特異度、識別率である。ここで感度とは、抗癌剤と免疫療法で併用効果があった患者を正しく効果ありと予測できた割合を意味する。一方、特異度とは、抗癌剤と免疫療法で併用効果のなかった患者を正しく効果なしと予測できた割合とする。

3.4.2 実験結果と考察

訓練サンプルとテストサンプルそれぞれに対する識別率を図 3.10 に示す。

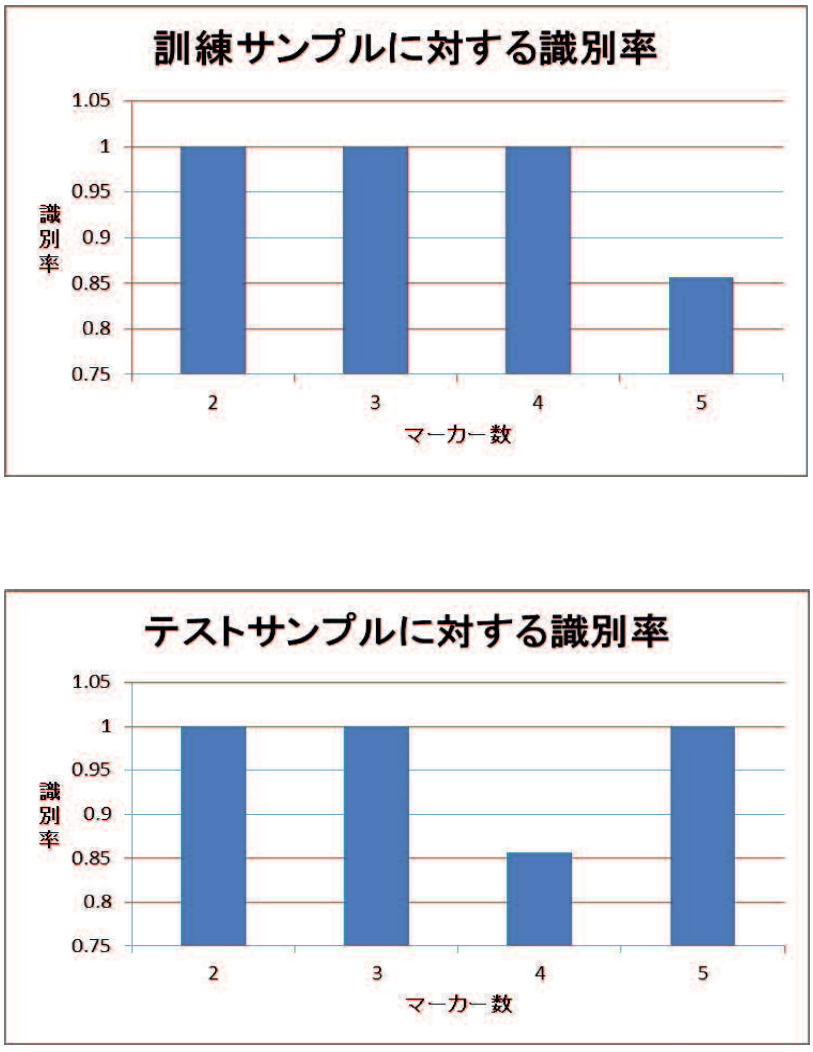


図 3.10　マーカー数と識別率の関係

表 3.9 にマーカー数毎に選択されたマーカーの組、表 2 にマーカーの組として選択されたマイクロ RNA の選択回数を示す。

表 3.9　テストサンプルに対する識別性能

マーカー数	マーカー(miR)					感度	特異度	識別率	8訓練サンプル上での事後確率の差分の絶対値	訓練サンプル数の正識別数
2	125a-3p	423-3p				1.00	1.00	1.00	0.97	8
	125a-3p	4419a				1.00	1.00	1.00	0.97	8
	125a-3p	4640-5p				1.00	1.00	1.00	0.97	8
	125a-3p	6765-5p				1.00	1.00	1.00	0.97	8
3	125a-3p	6875-5p	423-3p			1.00	1.00	1.00	1.68	8
	4433b-3p	125a-3p	423-3p			1.00	1.00	1.00	1.68	8
	4284	125a-3p	6765-5p			1.00	1.00	1.00	1.68	8
	4284	125a-3p	6089			1.00	1.00	1.00	1.68	8
	6892-5p	125a-3p	6765-5p			1.00	1.00	1.00	1.68	8
4	125a-3p	423-3p	6890-5p	6765-5p		1.00	0.67	0.86	2.47	8
	125a-3p	6887-5p	423-3p	6890-5p		0.75	1.00	0.86	2.47	8
	125a-3p	6887-5p	423-3p	3620-5p		0.75	1.00	0.86	2.47	8
	125a-3p	6887-5p	423-3p	6827-5p		0.75	1.00	0.86	2.47	8
	125a-3p	423-3p	4419a	3620-5p		0.75	1.00	0.86	2.47	8
	125a-3p	423-3p	4419a	6827-5p		0.75	1.00	0.86	2.47	8
	125a-3p	6887-5p	4419a	3620-5p		0.75	1.00	0.86	2.47	8
	125a-3p	6887-5p	4419a	6827-5p		0.75	1.00	0.86	2.47	8
5	1296-3p	887-5p	125a-3p	423-3p	6765-5p	1.00	1.00	1.00	2.73	7
	1296-3p	887-5p	125a-3p	423-3p	6089	1.00	1.00	1.00	2.73	7
	6892-5p	125a-3p	6875-5p	423-3p	6765-5p	1.00	1.00	1.00	2.73	7
	6892-5p	125a-3p	6875-5p	423-3p	6089	1.00	1.00	1.00	2.73	7
	1296-3p	887-5p	125a-3p	423-3p	4640-5p	1.00	1.00	1.00	2.73	7
	6892-5p	125a-3p	6875-5p	423-3p	4640-5p	1.00	1.00	1.00	2.73	7
	1296-3p	887-5p	125a-3p	423-3p	4419a	1.00	1.00	1.00	2.73	7
	6892-5p	125a-3p	6875-5p	423-3p	4419a	1.00	1.00	1.00	2.73	7
	4284	125a-3p	6875-5p	423-3p	6086	1.00	1.00	1.00	2.73	7
	1296-3p	887-5p	125a-3p	423-3p	6086	1.00	1.00	1.00	2.73	7
	6892-5p	125a-3p	6875-5p	423-3p	6086	1.00	1.00	1.00	2.73	7
	4284	4433b-3p	125a-3p	423-3p	6765-5p	1.00	1.00	1.00	2.73	7
	4284	4433b-3p	125a-3p	423-3p	6089	1.00	1.00	1.00	2.73	7
	4284	125a-3p	6875-5p	423-3p	6765-5p	1.00	1.00	1.00	2.73	7
	4284	125a-3p	6875-5p	423-3p	6089	1.00	1.00	1.00	2.73	7
	4284	4433b-3p	125a-3p	6887-5p	4419a	1.00	1.00	1.00	2.73	7
	4284	4433b-3p	125a-3p	423-3p	4640-5p	1.00	1.00	1.00	2.73	7
	4284	4433b-3p	125a-3p	423-3p	4419a	1.00	1.00	1.00	2.73	7
	4284	125a-3p	6875-5p	6887-5p	4419a	1.00	1.00	1.00	2.73	7
	1296-3p	887-5p	125a-3p	6887-5p	4419a	1.00	1.00	1.00	2.73	7
	6892-5p	125a-3p	6875-5p	6887-5p	4419a	1.00	1.00	1.00	2.73	7
	4284	125a-3p	6875-5p	423-3p	4640-5p	1.00	1.00	1.00	2.73	7
	4633-5p	3622b-5p	125a-3p	4419a	6089	1.00	1.00	1.00	2.73	7
	4284	4433b-3p	125a-3p	423-3p	6086	1.00	1.00	1.00	2.73	7
	4633-5p	3622b-5p	125a-3p	423-3p	6765-5p	1.00	1.00	1.00	2.73	7
	4633-5p	3622b-5p	125a-3p	423-3p	6089	1.00	1.00	1.00	2.73	7
	4284	125a-3p	6875-5p	423-3p	4419a	1.00	1.00	1.00	2.73	7
	4284	4433b-3p	125a-3p	4419a	6089	1.00	1.00	1.00	2.73	7
	4633-5p	3622b-5p	125a-3p	6887-5p	4419a	1.00	1.00	1.00	2.73	7
	4633-5p	3622b-5p	125a-3p	423-3p	4640-5p	1.00	1.00	1.00	2.73	7
	4633-5p	3622b-5p	125a-3p	423-3p	4419a	1.00	1.00	1.00	2.73	7
	4284	125a-3p	6875-5p	4419a	6089	1.00	1.00	1.00	2.73	7
	1296-3p	887-5p	125a-3p	4419a	6089	1.00	1.00	1.00	2.73	7
	6892-5p	125a-3p	6875-5p	4419a	6089	1.00	1.00	1.00	2.73	7
	4633-5p	3622b-5p	125a-3p	423-3p	6086	1.00	1.00	1.00	2.73	7

図 3.11 にマーカー数を２から３へ増加させた場合におけるテストサンプルに対する識別状況の可視化を示す。

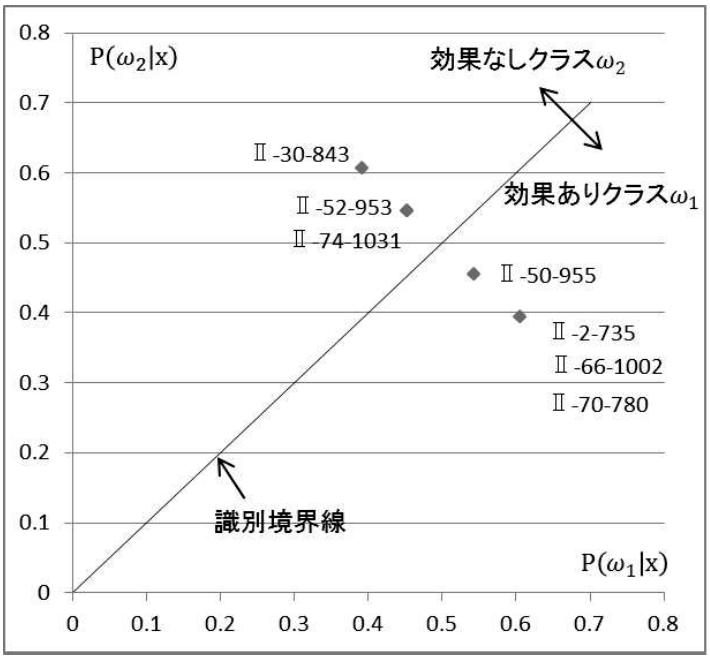


図 3.11 a　２マーカー（125a-3p、423-3p）を用いた場合

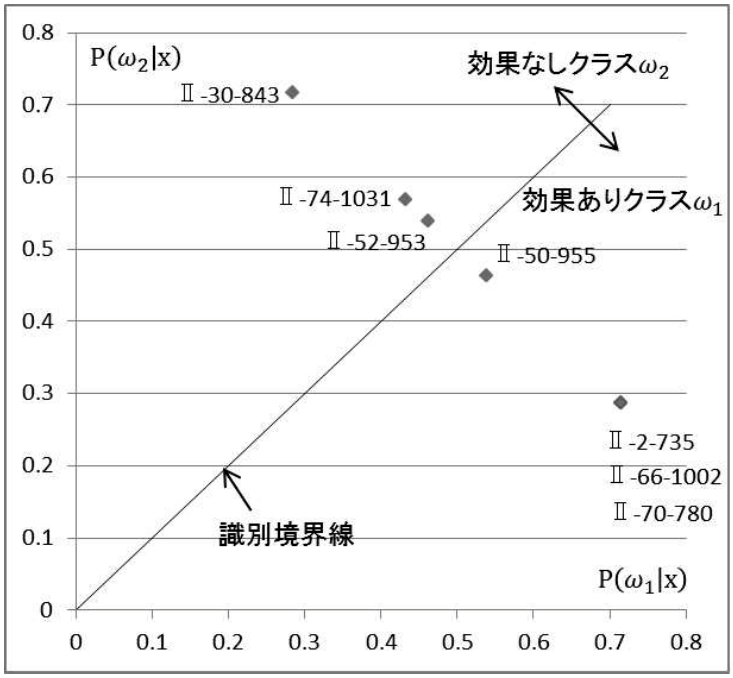


図 3.11 b　３マーカー（125a-3p、423-3p、6875-5p）を用いた場合

縦軸と横軸は共に事後確率とし、識別規則は事後確率の最大によることから、識別境界線は原点を通る 45° の直線となり、この直線の上側が効果なしクラス、下側が効果ありクラスと識別される。また、テストサンプルの各クラスの事後確率の組で定まる点からこの直線までの垂直距離 L の値が大きい程、識別が確実に行われていると解釈される。上図と下図の比較は、マーカーの組 125a-3p と 423-3p に 6875-5p を加えた場合の効果を示している。マーカーを追加することで、テストサンプルが識別境界線から離れていく様子が見られる。

図 3.10 より、訓練サンプルとテストサンプルそれぞれに対する識別率は、マーカー数 2 と 3 の場合が共に 1.00、つまり感度、特異度がともに 100%で、効果予測に適していると考えられる。テストサンプルに対する識別性能は表 3.9 に示しており、表 3.9 の中で頻出している 125a-3p と 423-3p、またそれらに 6875-5p を加えたマーカーの組を用いたテストサンプルに対する識別状況を可視化した結果を図 3.11 に示している。これより、マーカー数が 2 の 125a-3p と 423-3p を用いた場合よりもこれらに 6875-5p を加えた 3 マーカーの方が、テストサンプルが識別境界線から離れていることが確認された。これは、3 マーカーの識別性能が良いことを示している。

3.5 漢方薬の処方

漢方医学は医師の 9 割が用いていて、今日では西洋医学ともに現代医学には欠かせないものとなっている。漢方医学では、数々の生薬を調合して一人ひとり異なる内容で服用する。これは個別化医療そのものである。特に漢方医学の薬は「内側から治す」、「病気を未然に防ぐ」を基に、免疫力の向上を主に考えている。一方西洋医学における薬は、痛み止めや睡眠薬など即効性を重視している。このように漢方医学の考えは「根本的に内側から治す」を主体に、西洋医学の考えは「悪いところを切る」が根底にあり、両者の考え方は大きく異なる。それにも関わらず、それぞれに良い点と悪い点がある。

漢方医学は根本治療を目的としているため、即効性は乏しいが、副作用は少なく体全体のことを考えた治療と言える。WHO においても漢方医学の症状が疾病分類に登録され、漢方医学の国際化が進んでいる[16]、[17]。しかし、医学生や研修医、あるいは漢方を初めて用いようとする臨床医が、非常に多くの種類や成分、生薬を含む漢方薬から処方すべき漢方薬を選ぶことは容易ではない。そもそも大学では漢方医学を学ぶが、それは初級レベルの程度である。一方、臨床の現場に目を向けると、医師の経験や勘による処方には個人差があり、その処方には不確実性がある。

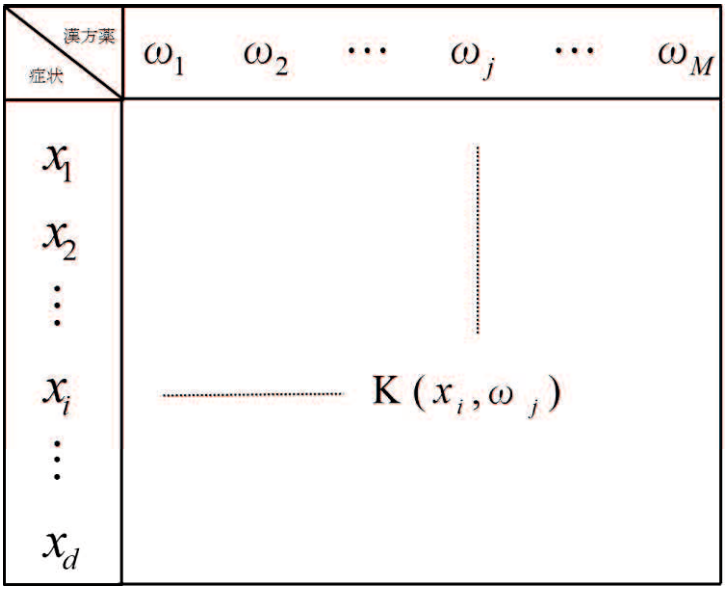
不確実性に対処する一つのアプローチに、統計的パターン認識[18]がある。統計的パターン認識では、不確実性を数値化して不確実な知識を利用できる知識に変換することができる[19]。本論文では、漢方薬の処方履歴データに基づいて、処方の不確実性を事後確率として数値化する。そして、事後確率を用いた離散 Bayes 識別則によって漢方薬の処方を考える。

3.5.1 離散 Bayes 識別則の修正

漢方薬処方データに用いるために、離散 Bayes 識別則を少し修正する。患者が$x_{i_1}, x_{i_2}, \cdots, x_{i_d}$という d 個の症状を訴えたとき、質的データである症状を成分とするパターンベクトル$\mathbf{x} = \left[x_{i_1}, x_{i_2}, \cdots, x_{i_d}\right]^T$で患者を表す。本手法では漢方薬$\omega_j$を処方する確率である事後確率$\mathrm{P}(\omega_j \mid \mathbf{x})$を求め、これに基づいて処方する。この事後確率$\mathrm{P}(\omega_j \mid \mathbf{x})$は、条件付き確率$\mathrm{P}(\mathbf{x} \mid \omega_j)$と$\mathrm{P}(\mathbf{x} \mid \overline{\omega_j})$、事前確率$\mathrm{P}(\omega_j)$と$\mathrm{P}(\overline{\omega_j})$から求められる。

まず事前確率と条件付き確率を定義する．いま d 個の症状があり、それに対して M 個の漢方薬が用意されている。症状x_iに対して漢方薬y_jを処方するとき、医師による評価値$K(x_i, \omega_j)$が与えられているとする。この$K(x_i, \omega_j)$を処方履歴データと呼び，これを表 3.10 に示す。

表 3.10　処方履歴データ



漢方薬 ω_s の事前確率 P(ω_s) と漢方薬 ω_s を処方するという条件のもとで症状 x_k が起こる条件付き確率 P(x_k | ω_s) を、それぞれ以下のように定める。まず漢方薬 ω_s の事前確率 P(ω_s) は

$$\mathrm{P}(\ \omega_s\ ) = \frac{\sum_{i=1}^{n} K(\ x_i\ ,\omega_s\ )}{\sum_{j=1}^{M}\sum_{i=1}^{n} K\ (x_i\ ,\omega_j\ )} \tag{1}$$

で与えられる。

一方、この漢方薬 ω_s を処方しないという事前確率 P($\overline{\omega_s}$) は

$$\mathrm{P}(\ \overline{\omega_s}\ ) = 1 - \ \mathrm{P}(\ \omega_s\ ) \tag{2}$$

で与えられる。

次に漢方薬 ω_s を処方するという条件のもとで、症状 x_k の条件付き確率 P(x_k | ω_s) は

$$\mathrm{P}(\ x_k \mid \omega_s\ ) = \frac{K(\ x_k\ ,\omega_s\ )}{\sum_{i=1}^{n} K(\ x_i\ ,\omega_s\ )} \tag{3}$$

で与えられる。一方、漢方薬 ω_s を処方しないという条件のもとで、症状 x_k の条件付き確率 P(x_k | $\overline{\omega_s}$) は

$$\mathrm{P}(x_k \mid \overline{\omega_s}) = \frac{\sum_{j=1}^{M} K(x_k, \omega_j) - K(x_k, \omega_s)}{\sum_{j=1}^{M}\sum_{i=1}^{d} K(x_i, \omega_j) - \sum_{i=1}^{d} K(x_i, \omega_s)} \tag{4}$$

で与えられる。

以上の準備のもと、事後確率 $\mathrm{P}(\omega_k \mid \mathbf{x})$ を求めることにする。一般に確率 $\mathrm{P}(\omega_k, \mathbf{x})$は、前述の仮定と乗法定理から

$$\mathrm{P}(\omega_k, \mathbf{x}) = \prod_{j=1}^{d} P\left(x_{i_j} \mid \omega_k\right) P(\omega_k) \tag{5}$$

と式変形される。

一方 $\mathrm{P}(\mathbf{x})$ は、標本空間が漢方薬ω_k を処方する事象(ω_k)と漢方薬 ω_k を処方しない事象($\overline{\omega_k}$)という二つの排反事象で構成されるため、

$$\mathrm{P}(\mathbf{x}) = \mathrm{P}(\omega_k, \mathbf{x}) + \mathrm{P}(\overline{\omega_k}, \mathbf{x})$$

となり，この式に式(1)を代入することにより

$$\mathrm{P}(\mathbf{x}) = \prod_{j=1}^{d} P\left(x_{i_j} \mid \omega_k\right) P(\omega_k) + \prod_{j=1}^{d} P\left(x_{i_j} \mid \overline{\omega_k}\right) P(\overline{\omega_k}) \tag{6}$$

で与えられる。

また確率 $\mathrm{P}(\omega_k, \mathbf{x})$ は

$$\mathrm{P}(\omega_k, \mathbf{x}) = P(\mathbf{x}) P(\omega_k \mid \mathbf{x}) \tag{7}$$

と表現されることもある。式 (5),(6)と(7)から事後確率 $P(\omega_k \mid \mathbf{x})$ は

$$P(\omega_k \mid \mathbf{x}) = \frac{\mathrm{P}(\omega_k, \mathbf{x})}{P(\mathbf{x})}$$

$$= \frac{\prod_{j=1}^{d} P\left(x_{i_j} \mid \omega_k\right) P(\omega_k)}{\prod_{j=1}^{d} P\left(x_{i_j} \mid \omega_k\right) P(\omega_k) + \prod_{j=1}^{d} P\left(x_{i_j} \mid \overline{\omega_k}\right) P(\overline{\omega_k})} \tag{8}$$

となる。つまり、事後確率 $P(\omega_k \mid \mathbf{x})$ は，事前確率 $P(\omega_k)$ と $P(\overline{\omega_k})$ 及び条件付き確率 $P\left(x_{i_j} \mid \omega_k\right)$ と $P\left(x_{i_j} \mid \overline{\omega_k}\right)$ から求めることができる。

求められた事後確率は、症状のパターン $\mathbf{x}$ に対して漢方薬ω_k を処方すべきか

否かという識別に用いられる。事後確率 $P(\omega_k \mid \mathbf{x})$ とは、症状のパターン $\mathbf{x}$ という条件のもとで、漢方薬 ω_k を処方する確率を意味する。この事後確率が大きい程, 症状のパターン $\mathbf{x}$ に対して漢方薬 y を処方するという妥当性が高くなる。従って、この症状の患者に対しては統計的パターン認識の Bayes 識別則に従って、事後確率が最大となる漢方薬を処方する。すなわち患者が症状 $\mathbf{x}$ を訴えるとき、

$$\mathrm{P}(\widehat{\omega_k} \mid \mathbf{x}) = \max_{\omega_k} P(\omega_k \mid \mathbf{x})$$

ならば、事後確率が最大の漢方薬 $\widehat{\omega_k}$ を処方する。事後確率が最大とは、過去の処方履歴のデータから多くの医師が最も妥当であると評価していることを意味する。以上の流れを図 3.12 に示す。

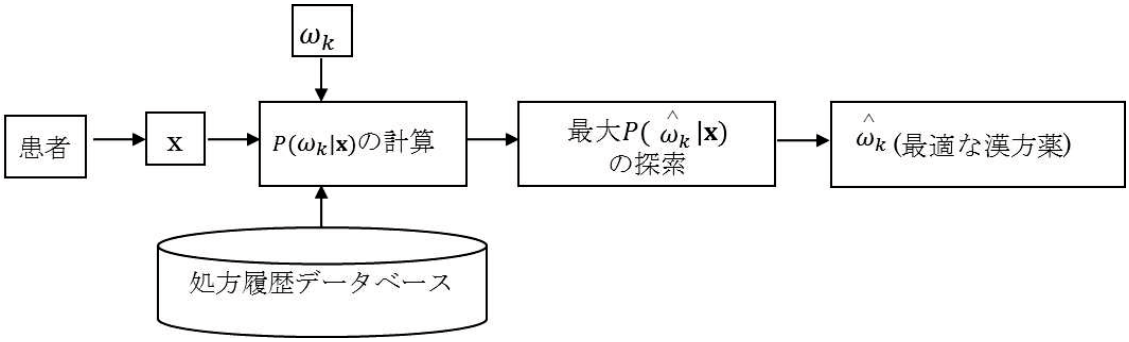


図 3.12　最適な漢方薬処方の流れ

3.5.2 実験方法

本論文で用いる処方履歴データは、次のように求められた。専門の異なる漢方専門医 27 名に対してアンケート形式で、患者がある症状を訴えたとき、医師がその症状に対して処方している漢方薬を七段階の数値で評価したスコア値を集計した。そして 27 名から得られたスコア値の平均値を、ある症状に対する漢方薬の評価値とした。本論文では、55 症状と 128 個の漢方薬を対象とする。

本手法では患者が訴えた複数の症状に対する事前確率と条件付き確率を処方履歴データから計算した後、それらを用いて事後確率を計算する。事後確率が最大となる漢方薬を最適として医師に提示する。本実験では、漢方薬の典型的な処方として症状の組を、「月経不順」と「月経困難症」、「めまい」と「冷え」、「頭痛」と「悪寒」の３例とする。

3.5.3 実験結果と考察

各例に対する事後確率の１位から３位までの漢方薬を表 3.11 から表 3.13 に示す。

表 3.11　症状「月経不順」と「月経困難症」の場合

順位	漢方薬(in Japanese)	事後確率
１位	桂枝茯苓丸（Keishibukuryougan）	0.75
２位	当帰芍薬散（Toukisyakuyakusan）	0.69
３位	加味逍遥散（Kamisyouyousan）	0.51

表 3.12　症状「めまい」と「冷え」の場合

順位	漢方薬(in Japanese)	事後確率
１位	真武湯（Shinbutou）	0.66
２位	当帰芍薬散（Toukisyakuyakusan）	0.14
３位	半夏白朮天麻湯（Hangebyakujututenmatou）	0.11

表 3.13　症状「頭痛」と「悪寒」の場合

順位	漢方薬(in Japanese)	事後確率
1 位	葛根湯（Kakkontou）	0.69
2 位	麻黄湯（Maoutou）	0.34
3 位	麻黄附子細辛湯（Maoubushisaishintou）	0.29

症状が「月経不順」と「月経困難症」のとき、本手法では事後確率が最大の 75%である桂枝茯苓丸（Keishibukuryougan）を最適として選択した。症状が「めまい」と「冷え」のとき、本手法では事後確率が最大の 66%である真武湯（Shinbutou）を最適として選択した。症状が「頭痛」と「悪寒」のとき、本手法では事後確率が最大の 69%である葛根湯（Kakkontou）を最適として選択した。

本実験で示した処理を Web システム上で実装した。図 3.13 に漢方薬検索画面を示す。検索条件である「症状の入力」は最大 3 つまで選択することができる。なお少なくとも 1 つは必ず選択することが必要である。

図 3.13　漢方薬の検索画面

漢方薬はセレクトボックスの隣にある番号順、５０音順のそれぞれボタンをクリックすることで表示の順番を変更することができる。次に「漢方薬の選択」は任意であり、選択した場合は、選択した漢方薬の処方確率が表示される。検索条件の選択が完了してからは「検索」をクリックする。選択した症状に対して，処方確率が大きい順に 10 位までの漢方薬とその処方確率、及び選択した漢方薬の処方確率を表示する。図 3.14 に検索結果の表示画面を示す。ここでは、症状に疲労・全身倦怠感を選択し、漢方薬に葛根湯を選択した。その結果、葛根湯の処方確率は 0%となり、疲労・全身倦怠感を訴えている患者への処方に適していないことが示された。選択した漢方薬の下部に推奨される漢方薬を大きい順に 10 位まで表示しており、疲労・全身倦怠感には補中益気湯が一番適していることが確認できる。なお、ここで表示する漢方薬は、処方確率が 1%以上の漢方薬のみである。

図 3.14　検索結果の表示画面

本 Web システムを用いることによって，医学生や研修医，あるいは漢方を初めて用いようとする臨床医を対象として，漢方薬の処方履歴データを用いた統計的パターン認識に基づいて漢方診療や教育の支援を行うことができる.

3.6 おわりに

本論文では、離散 Bayes 識別則を提案し、4 つの個別化医療問題に適用した。その結果についての検討を以下に示す。

初めに、本論文では、離散 Bayes 識別則を難治性の高い肝癌の早期再発を高精度で予測問題に適用した。離散 Bayes 識別則は、質的データだけでなく量的データを離散化することで、全てのマーカーを取り扱うことができる。肝癌再発予測の実験を行い、既存の代表的なスコア式である Tokyo Score, Modified JIS, TNM 分類よりも高精度で肝癌の早期再発を予測することができた。離散 Bayes 識別則により個別化医療を実現することが期待される。本手法では、計算を簡単にするため、特に議論することなく独立性の仮定を用いた。独立性を検証することは困難であるが、Bayesian networks[11]を通して独立性に関する検討は今後の課題となる。また本手法の識別性能は、マーカー値を離散化するときに用いるカットオフ値に依存するため、その最適化も興味ある課題の一つである。

次に、早期胃癌のリンパ節転移の予測問題では、深達度、ly と v がマーカーとして有用であることを示した。具体的に言えば、深達度が M か SM1、ly が陰性、v が陰性であるときはリンパ節転移はないと考えられ、それぞれのマーカーが一つでも上記と異なる場合、リンパ節転移の可能性があると考えられる。識別性能は、感度 1.00,特異度 0.86 であり、高い感度を保ったまま現実的な特異度

を維持している。言い換えると、特異度が高いため、リンパ節転移がなく、それゆえ必要のない手術を行うリスクを低減することが期待される。しかし、本論文で使用したリンパ節転移が存在する症例は 9 サンプルと極めて少なく、性能の評価が制限されている。

次に、離散 Bayes 識別則による大腸癌における抗癌剤と免疫療法の併用効果の予測では、 感度、特異度ともに 100%を達成する３マーカーの組（125a-3p と 423-3p、6875-5p）を同定した。免疫治療と用いる抗癌剤の効果を投与前に正確に予測することで、患者個々に最適な治療を提供できる。つまり、個別化医療の実現に繋がると考えられる。この実験の問題点も、用いたサンプル数が１５と、非常に少ないことである。

最後に、離散 Bayes 識別則を漢方薬の処方問題へ適用した。患者が訴える症状に対して本手法で求めた最適な漢方薬は、３例全てにおいて医師の処方と一致しており、本手法は妥当であるといえる。今後、漢方専門医の更なる評価を受けて本手法の実用性を検討する。

参考文献

1） N. Iizuka, Y. Hamamoto, R. Tsunedomi, M. Oka, Translational microarray systems for outcome prediction of hepatocellular carcinoma, Cancer Science, Vol.99, pp.659-665, 2008.

2) N. Iizuka, M. Oka, Y. Hamamoto et al., Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, Lancet, Vol.361, pp.923-929, 2003.

3) A.K. Jain, R.W. Duin, and J. Mao, Statistical pattern recognition: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.1, pp.4-37, 2000.

4) R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, Second Edition, John Wiley & Sons, 2001.

5) A.K. Jain, R.C. Dubes and C.-C. Chen, Bootstrap techniques for error estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.9, No.5, pp.628-633, 1987.

6) Y. Tokumitsu, N. Iizuka, H. Ogihara, Y. Hamamoto et al., An Accurate Prognostic Staging System for Hepatocellular Carcinoma Patients after Curative Hepatectomy, International Journal of Oncology, Vol.46, pp.944-952, 2015.

7) A. Webb, Statistical Pattern Recognition, Second Edition, John Wiley & Sons, 2002.

8) W.J. Youden, Index for Rating Diagnostic Tests, Cancer, Vol.3, pp.32-35, 1950.

9) C.J. van Rijsbergen, Foundation of evaluation, Journal of Documentation, 30, pp.365-373, 1974.

10) D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation, Journal of Machine Learning Technologies, Vol.2, No.1, pp.37-63, 2011.

11) J. Pearl, Bayesian Networks: a Model of Self-Activated Memory for Evidential Reasoning, Proceedings, Cognitive Science Society, UC Irvine, pp.329-334, 1985.

12) 最新がん統計：[国立がん研究センター　がん登録・統計]」, http://ganjoho.jp/reg_stat/statistics/stat/summary.html.

13) A.K. Jain, R.W. Duin, and J. Mao, Statistical pattern recognition: A review, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, No.1, pp.4-37, 2000.

14) 浜本義彦, 統計的パターン認識入門, 森北出版, 2009.

15) がん免疫療法：基礎研究から臨床応用にむけて, http://leading.lifesciencedb.jp/4-e005/, DOI: 10.7875/leading.author.4.e005.

16) WHO Medicines. Traditional and Complementary Medicine, http://www.who.int/medicines/areas/traditional/en/.

17) ICD-11 Beta Draft, http://apps.who.int/classifications/icd11/browse/f/en.

18) R.O. Duda, P.E. Hart and D.G. Stork, Pattern Classification, Second Edition, John Willey & Sons, New York, 2001.

19) C.R. Rao, Statistics and Truth: Putting Chance to Work, 2nd Edition, CSIR, New Delhi, 1987.

第 4 章　結論

4.1 まとめ

本論文では、初めに機械学習による個別化医療の問題点として、質的データを用いることができない点を指摘した。この問題点の解決策として、質的データも扱える離散 Bayes 識別則を提案し、それにより入力データに対する制約を解消した。更に、離散 Bayes 識別則を、具体的な個別化医療問題に適用し、その有用性を検討した。以下に各章を要約し、本論文のまとめとする。

第 2 章では、統計的パターン認識について概説し、個別化医療問題を統計的パターン認識問題として定式化した。その後、離散 Bayes 識別則を提案し、その評価方法も定めた。

第 3 章では、まず離散 Bayes 識別則を用いて肝癌の早期再発の予測問題を取り扱った。その結果、離散 Bayes 識別則に基づく独自の特徴選択により、識別に有用な 4 標的マーカー（腫瘍数×腫瘍サイズ、vp、ICG、Liver damage）を同定し、感度 0.86、特異度 0.49 を達成した。実験として、訓練サンプル数の影響の調査、識別の計算コストの調査、他手法との予測精度の比較を行った。初めに、訓練サンプル数の識別性能に与える影響を調べた。訓練サンプル数を増加させるにつれて識別性能は収束していく様子が観察され、訓練サンプル数は十分であることが確認できた。次に、識別に要する計算コストについて調べ、識別に要する時間が線形のオーダーであることが確認された。このことから、大量の計算が必要とされるビッグデータに対しても計算コストがそれほど増加せず、通常のコンピュータで実行可能であることが示せた。次に、離散 Bayes 識別則の予測性能を従来の肝臓のスコア式による予測性能と比較した。早期再発の予測問題であるため、予測性能は一定の特異度を保ちつつ高感度であるこ

とが求められる。実験では、提案手法の予測性能は既存のスコア式ほど特異度を低下させることなく、早期再発の予測に重要な指標である感度が高いことを示しており、提案手法は医学的に有効であるという結果を得た。

次に、離散 Bayes 識別則を用いた早期胃癌のリンパ節転移の予測問題に取り組んだ。この解析では、早期胃癌に対する ESD 治療後の特異度が低く、必要としない手術が行われていることが問題であった。leave one out 法と再代入法を組み合わせた特徴選択により、診断・予測に有用な 3 標的マーカー（深達度、ly、v）を同定した。最適マーカーの組合せを用いた離散 Bayes 識別により感度 1.00、特異度 0.86 を達成した。感度 1.00 で転移見逃しの問題を解決しており、提案手法の有用性を示せた。

更に、大腸癌における抗癌剤と免疫療法の併用効果の予測問題に取り組んだ。抗癌剤と免疫療法を併用し、効果のある患者と効果のない患者を層別化するマーカーを離散 Bayes 識別則で同定した。実験結果より、抗癌剤と免疫療法の併用効果を予測するためのマーカーとして、125a-3p、423-3p、6875-5p の 3 標的マーカーを同定し、予測性能として感度と特異度共に 1.00 を達成した。

最後に、離散 Bayes 識別則を応用した漢方薬の処方問題に取り組んだ。漢方薬の処方は、医師の経験や勘によるため、処方には不確実性があった。この不確実性を数値化して漢方薬の処方に役立てる試みである。離散 Bayes 識別則を用いた処方問合せにより、異なる症状の組合せ 3 例に対して全て医師の処方と一致した。これにより、本手法の妥当性が示された。以上が本論文における成果である。以下、今後の研究の課題と展望を述べ、本論文のむすびとする。

本論文では、提案手法を様々な個別化医療問題へ適用したが、識別器の設計、評価に用いたサンプル数はいずれも少数で、しかも単一施設のものであった。今後他施設からのデータも用いて、離散 Bayes 識別則を評価することが課題の

一つである。

上記の他に今後の課題として、マーカーのカットオフ値の最適化問題がある。これまで、量的データを質的データに変換する際、マーカーに対するカットオフ値は医学的に医師によって予め定められていた。このマーカーが、例えば遺伝子、タンパク質など分子標的であった場合、医師がカットオフ値を定めるのは一般に困難である。そのためコンピュータを用いた最適化手法によりカットオフ値の最適化が必要となる。

今後の展望としては、創薬[1]に対して離散 Bayes 識別則を活用したいと考えている。現在、新薬の開発から完成に至るまでにほとんどの場合で１０年以上の歳月と莫大な費用を要する。この期間の中で最も時間と費用のかかるプロセスは治験（臨床試験）[2]である。治験は開発中の新薬が人間に対して安全なものであるかを調べるため、健常者または患者を対象に行っている試験である。その治験対象者は開発中の新薬に対して効果があるか否かは事前に分からないため、一般に数多くの患者と健常者を必要としている。このとき、事前に薬の効果の有無を予測できれば薬の効く患者のみに対象者を選別できる。これはパターン認識問題における患者の層別化である。患者の層別化には標的マーカーの選択と識別がある。例えば、薬の効果の有無を離散 Bayes 識別則により予測することで、効果のある患者と効果の無い患者を識別することができる。これにより、効果のあるクラスにのみ治験の対象者を絞ることが可能となるため、治験サイズを縮小し、費用と時間の大幅な削減に繋がるメリットがある。このことから提案手法は創薬に貢献できると期待している。

参考文献

1）創薬と治験 | 公益社団法人日本薬学会,

http://www.pharm.or.jp/souyaku/cro_2.shtml.

2）治験 | 厚生労働省,

http://www.mhlw.go.jp/stf/seisakunitsuite/bunya/chiken.html.

謝辞

本論文をまとめる現在に至るまで、終始懇切丁寧なご指導を頂きました山口大学大学院創成科学研究科教授 浜本義彦先生に心より感謝の意を表しますと共に厚く御礼申し上げます。また、本論文をまとめるにあたり、有益な御教示を頂きました山口大学大学院創成科学研究科教授 松藤信哉先生、山口大学大学院創成科学研究科教授 山口真悟先生、山口大学大学院創成科学研究科准教授 長篤志先生、山口大学大学院創成科学研究科准教授 藤田悠介先生に深く謝意を表します。

また、漢方薬の処方の研究において、様々な御助言及び御協力を頂きました広島大学大学院医歯薬保健学研究科教授 飯塚徳男先生に深く感謝いたします。

早期胃癌におけるリンパ節転移の予測の研究では、様々な御助言及び御協力を頂きました山口大学大学院医学系研究科教授 西川潤先生、山口大学大学院医学系研究科助教 五嶋敦史先生に深く感謝いたします。

また、肝癌の早期再発予測の研究において、様々な御助言及び御協力を頂きました山口大学大学院医学系研究科助教 徳光幸生先生に深く感謝いたします。

大腸癌における抗癌剤と免疫療法の併用効果の予測の研究では、様々な御助言及び御協力を頂きました山口大学大学院医学系研究科教授 硲彰一先生、山口大学大学院医学系研究科助教 恒富亮一先生に深く感謝いたします。

最後に、大学進学、大学院進学に至るまで肉体的、精神的に支えていただきました、祖父母、両親、妹に感謝の意を表します。

付録

・付録 1　離散 Bayes 識別則の注意すべき点

離散 Bayes 識別則では、選択した複数のマーカーそれぞれの分割におけるサンプルの割合が両クラスで同数である場合、識別ができないという弱点がある。付表 1.1 の数字を用いて具体例を示す。

付表 1.1　各マーカーにおける分割と患者数

マーカーの分割	ω_1	ω_2
$x_{1(1)}$	1	10
$x_{1(2)}$	9	90
$x_{2(1)}$	3	30
$x_{2(2)}$	5	50
$x_{2(3)}$	2	20

クラスω_1：10 サンプル、クラスω_2：100 サンプルで、マーカーx_1とx_2が対象となっており、ある患者の測定値が分割$x_{1(1)}$と$x_{2(3)}$に属したとする。

このとき、クラスω_1に対する条件付き確率は、

$$\mathrm{P}\left(x_{1(1)}\middle|\omega_1\right)=\frac{1}{1+2}=\frac{1}{3}$$

$$\mathrm{P}\left(x_{2(3)}\middle|\omega_1\right)=\frac{2}{1+2}=\frac{2}{3}$$

である。一方、クラスω_2に対する条件付き確率は、

$$\mathrm{P}\left(x_{1(1)}\middle|\omega_2\right)=\frac{10}{10+20}=\frac{1}{3}$$

$$\mathrm{P}\left(x_{2(3)}\middle|\omega_2\right)=\frac{20}{10+20}=\frac{2}{3}$$

となる。よって、最終的に事後確率はどちらも 0.5 となり識別ができない。

しかし、３マーカー以上ではこのようなケースは稀である。それにも関わらず両クラスで分割されたサンプル数の割合が接近している場合、これは注意すべき問題であるといえる。