生体医工学
43(3)：416–423, 2005

研　究

統計的識別手法を用いた C 型肝炎患者の血液検査データによる病態識別

青木　一真*・黒 柳　奨*・岩 田　彰*・山内　一信**

Discriminating the Condition of Hepatitis C Using Blood Examination Data by Statistical Methods

Kazuma AOKI,* Susumu KUROYANAGI,* Akira IWATA,* Kazunobu YAMAUCHI**

Abstract Presently, doctors predict the condition of hepatitis C using blood examination data based on their professional experience, and patients are then diagnosed by performing a liver biopsy to obtain a definite diagnosis. However, liver biopsies are a high-risk procedure and can be troublesome. In this paper, we suggest a new method that is easier and more accurate. It uses the SVM (support vector machine), which is one of the most effective learning machines, and SFFS (sequential forward floating search), which is a feature selection. The combination of SVM and SFFS make it possible to eliminate the unnecessary examination of various items. It also helps to obtain high accuracy compared to using only SVM. Performance was drastically improved by applying our new method to the blood examination data for hepatitis C.

Keywords: support vector machine, feature selection, sequential forward floating search, hepatitis C.

1. はじめに

ウイルス性肝炎には A 型，B 型，C 型，D 型，E 型があり，日本では A 型，B 型，C 型が多数を占める．肝硬変や肝臓癌の原因となる肝炎ウイルスの約 8 割が C 型肝炎である．また，C 型急性肝炎患者の 6 ～ 8 割が慢性肝炎に移行すると言われ，30 年あまりかけて進行し肝硬変となる．そこで肝炎の状態に合わせて適切な治療をする必要がある[1–3]．

肝炎の病態は肝臓の線維化（fibrosis：F）によって定義することができ，線維化の状態を知ることのできる検査が肝生検である．線維化はその度合いによって，門脈域から線維化が進展し小葉が改築され肝硬変へ進展する段階を線維化なし（F0），門脈域の線維性拡大（F1），bridging fibrosis（F2），小葉のひずみを伴う bridging fibrosis（F3）までの 4 段階に区分する．さらに結節形成傾向が全体に認められる場合は肝硬変（F4）と分類する．

2005 年 3 月 8 日受付，2005 年 7 月 11 日改訂
Received March 8, 2005; revised July 11, 2005.
* 名古屋工業大学大学院工学研究科情報工学専攻
Department of Computer Engineering, Nagoya Institute of Technology
** 名古屋大学大学院医学系研究科医療管理情報学
Medical Information & Management Science, Nagoya University Graduate School of Medicine

肝炎の進行具合により適切な治療法があり，患者の病態を判別することはきわめて重要である．現在 C 型肝炎患者の検査として，主として行われる検査が血液検査である．血液検査の結果から病態を判別するには，医師がこれまでの経験に基づいて推定するが，確定診断を得るために肝生検と呼ばれる検査が行われている．肝生検は，細く長い針を皮膚の表面から直接肝臓に刺して，肝臓の組織を採取し，顕微鏡で肝臓の状態を調べる検査である．肝生検は一番正確かつ信用できる肝炎の診断法であるが，患者に侵襲があり，数日間の入院を要する場合がある．このように肝生検においても手間や危険性といった問題点がある．

これに対し，血液検査の結果から肝疾患の診断を支援するシステムの 1 つとして，小栗らにより NI-SYS が提案されている[4]．NI-SYS は専門医へのアンケートをもとに慢性活動性肝炎，肝硬変，肝細胞癌などの肝疾患の各疾患ごとに重要となる血液検査データを選別し，それぞれの疾患を専門的に診断するニューラルネットワークを作成，これらの結果を統合することにより診断をするシステムである．同論文の結果によると NI-SYS を用いることで同様のデータに対する医師へのアンケート調査結果（63%）よりも高い識別率（75.4%）を得ることができた．

NI-SYS においては，各種の血液検査結果より診断に必要な検査項目を医師の知見を元に選別することで識別率の向上をはかっている．よって血液検査からの疾患識別にお

いては，可能な限り多くの検査項目を用いるのではなく，むしろ検査項目を選別することの重要性が明らかとなった．しかし，医師の知見を元に検査項目を選別するのは非常な労力を必要とするため，なんらかの方法で自動的に検査項目を選別することが望ましい．

そこで本研究では，C 型肝炎の病態（線維化）識別に対して，有効な検査項目を自動的に選択（特徴選択）することで識別率の向上をはかる手法を検討し，統計的識別手法を用いて血液検査データから病態を判別するシステムを提案する．検査項目の選択については特徴選択の代表的な手法である SFFS（sequectial forward floating search）を用いることとした．

しかし，SFFS に代表される特徴選択手法は選択した特徴に対して学習器械を毎回構築し選択結果を評価する必要があるため，階層型のニューラルネットワークのように学習に多大な演算コストの必要となる手法は用いることができない．これに対して非常に性能の高い非線型識別器械として，近年 SVM（support vector machine）が注目を浴びている．SVM はその高い識別性能とともに学習が非常に高速であることが知られており，本研究の目的に即した手法であるといえる．ただし SVM は 2 クラス分類手法であり基本的に複数のクラスを分類することはできない．そこで，本研究では C 型肝炎の病態のうち F0 から F2 をクラス f1，F3 から F4 をクラス f2 とし，これら 2 つのクラス分類を血液検査データから行うことを目的とし，特徴選択 SFFS と識別器械 SVM を組み合わせたシステムの有効性を検証した．

なお，実験に用いた血液検査データは C 型肝炎患者の血液検査データで，それぞれのサンプルは各患者があらかじめ肝生検を受けることで線維化の段階がわかっているものである．

2.　SVM（Support Vector Machine）

近年，パターン認識の分野で Vapnik[5]によって提案された SVM が注目されている．SVM はニューラルネットワークや nearest neighbor などの現在知られている多くの手法の中でも，最も認識性能の優れた学習モデルの一つである．SVM は線形 SVM（linear support vector machine: LSVM）と非線形 SVM（nonlinear support vector machine: NSVM）に大別できる．ここではまず LSVM のハードマージンとソフトマージンについて説明をし，次に NSVM を説明する．

2･1　LSVM（ハードマージン）

与えられた l 個の学習データが超平面で誤りなく分離できる場合を考える．各々の学習データは，特徴ベクトル $\boldsymbol{x}_i \in \mathbf{R}^n (i=1,\cdots,l)$, それに割り当てられたクラス $y_i \in \{-1,1\}$ の組からなる．

正のサンプルと負のサンプルを分離する超平面（分離超平面）の方程式を

$$H_0 : \boldsymbol{\omega}\cdot\boldsymbol{x}+b=0$$

とする．ここで $\boldsymbol{\omega}$ は超平面の法線ベクトルで，b は定数項である．これらのパラメータを変更することで識別面をコントロールできる．

次に d_+, d_- を分離超平面から最も近い正，負のサンプルまでの最短距離とする．この最短距離を分離超平面のマージンと呼ぶ．線形分離可能な場合，SVM はマージンが最大である分離超平面を求める問題である．

すべての学習データは H_0 によって分離されるため，次の制約条件を満たさなければならない．

$$y_i(\boldsymbol{x}_i\cdot\boldsymbol{\omega}+b)-1\geq 0 \quad (i=1,\cdots,l) \tag{1}$$

学習データと超平面の距離は，

$$\frac{|\boldsymbol{x}_i\cdot\boldsymbol{\omega}+b|}{\|\boldsymbol{\omega}\|} \tag{2}$$

と表せる．

したがって，最大マージンを持つ超平面を求めるには，式（1）の制約条件のもとで $\|\boldsymbol{\omega}\|^2$ を最小化すればよい．以上より，SVM は次の制約付き最適化問題に定式化できる．

$$\text{目的関数：}\frac{1}{2}\|\boldsymbol{\omega}\|^2 \quad \rightarrow \quad \text{最小化} \tag{3}$$

$$\text{制約条件：} y_i(\boldsymbol{x}_i\cdot\boldsymbol{\omega}+b)-1\geq 0 \quad (i=1,\cdots,l) \tag{4}$$

一般に制約付きの問題は，ラグランジュの乗数法を用いると，より簡単な問題に帰着することが多い．この問題を解くためにラグランジュの乗数法を用いると次式のようになる．

$$\text{目的関数：}\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i\cdot\boldsymbol{x}_j \rightarrow \text{最大化} \tag{5}$$

$$\text{制約条件：}\alpha_i\geq 0 \quad (i=1,\cdots,l),\ \sum_{i=1}^{l}\alpha_i y_i=0 \tag{6}$$

この問題を数値計算で解くと，多くの α_i が 0 となり，$\alpha_i \neq 0$ を満たすものが最小距離のサンプル（サポートベクトル）に対応することが知られている．最適な α から $\boldsymbol{\omega}$ を得るには次式を用いる．

$$\boldsymbol{\omega}=\sum_{i=1}^{l}\alpha_i y_i \boldsymbol{x}_i \tag{7}$$

2･2　LSVM（ソフトマージン）

ハードマージンでは，学習サンプルは超平面によって完全に分離できると仮定している．パターン認識の実問題で線形分離可能な場合は稀であり，実際的な問題に SVM を使うには多少の識別誤りは許すように制約を緩める必要がある．これをソフトマージンと呼ぶ．

ソフトマージンでは，反対側にどのくらいは入り込んだかの距離を，スラック変数 $\xi_i \geq 0 \ (i=1,\cdots,l)$ を用いて，$\xi_i / \|\boldsymbol{\omega}\|$ と表す．また ξ_i の和，$\sum_{i=1}^{l}\xi_i$ はできるだけ小さい方が望ましい．これを線形分離可能な場合の問題の目的関数，式（3）に加えたものをこの問題の目的関数とする．ま

た制約条件を次のように緩める．

$$\text{目的関数：}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{l}\xi_i \rightarrow \text{最小化} \qquad (8)$$

$$\text{制約条件：}\ y_i(\boldsymbol{x}_i\cdot\boldsymbol{\omega}+b)\geq 1-\xi_i\ (i=1,\cdots,l) \qquad (9)$$

ここでCは，第1項のマージンの大きさと，第2項のはみ出しの程度とのバランスを決めるパラメータであり，設定は実験的に行う必要がある．

前節と同様にして，ラグランジュ乗数法を用いて定式化をすると次のようになる．

$$\text{目的関数：}\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i\cdot\boldsymbol{x}_j \rightarrow \text{最大化} \qquad (10)$$

$$\text{制約条件：}\ 0\leq\alpha_i\leq C\ (i=1,\cdots,l),\ \sum_{i=1}^{l}\alpha_i y_i=0 \qquad (11)$$

2・3　NSVM

ソフトマージンを用いたとしても，本質的に非線形で複雑な識別問題に対しては，良い識別率を得ることができるとは限らない．このような問題に対しての解決法として，特徴ベクトルを非線形変換して，その空間で線形分離を行うカーネルトリックと呼ばれる方法がある．

一般に，線形分離の可能性はサンプル数が大きくなるほど困難になるが，特徴空間ベクトルの次元が大きくなるほど容易になる．今，写像

$$\Phi:\mathbf{R}^n\mapsto\mathbf{R}^q$$

を用いて学習データをより高次元の空間$\mathbf{R}^q$に写して，その空間で線形識別を行うことを考える．

一般に，このような非線形写像によって変換した特徴空間の次元は非常に大きくなりがちになり，結果的に膨大な計算量が必要となる．しかしSVMでは目的関数や識別関数が入力パターンの内積のみに依存した形になっており，内積が計算できれば最適な識別関数を求めることが可能である．つまり，非線形に写像した空間で2つの要素$\Phi(\boldsymbol{x}_1)$, $\Phi(\boldsymbol{x}_2)$の内積が

$$\Phi(\boldsymbol{x}_1)\cdot\Phi(\boldsymbol{x}_2)=K(\boldsymbol{x}_1,\boldsymbol{x}_2) \qquad (12)$$

のように，入力特徴$\boldsymbol{x}_1,\boldsymbol{x}_2$のみから計算できるなら，非線形写像によって変換された特徴空間での特徴$\Phi(\boldsymbol{x}_1)$, $\Phi(\boldsymbol{x}_2)$を実際に計算する代わりに，$K(x_1,x_2)$から最適な非線形写像を求められる．このようなKのことをカーネルと呼ぶ．本研究では次に示すGaussカーネルを用いた．

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2)=\exp\left(-\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|^2}{2\sigma^2}\right) \qquad (13)$$

このような写像を使うと，識別関数は式(7)より，

$$f(\Phi(x))=\sum_{i=1}^{l}\alpha_i y_i K(\boldsymbol{x},\boldsymbol{x}_i)+b \qquad (14)$$

となる．同様に学習の問題も，

目的関数：

$$\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\rightarrow\text{最大化} \qquad (15)$$

制約条件：

$$0\leq\alpha_i\leq C\ (i=1,\cdots,l),\ \sum_{i=1}^{l}\alpha_i y_i=0 \qquad (16)$$

と書くことができる．

Gaussカーネルを用いたNSVMの学習においては，学習サンプル間のカーネル関数Kをすべての学習サンプルの組み合わせについて計算し，これらを用いてα_iを求める．しかし，これらカーネル関数は一度計算したものをルックアップテーブルに保管しておくことで再度計算をする必要がなくなるため，これにより高速な学習が可能となる．

学習後のSVMにおいて新たなデータを認識する場合には，$\alpha_i\neq 0$となった学習サンプル（すなわちサポートベクトル）すべてに関して，認識データとのカーネル関数値を計算し，式(14)により得られた計算結果の正負により識別を行う．

NSVMはマージン最大化を目的として学習を行うため，同じ非線型識別機械である三層階層型ニューラルネットワークなどと比較した場合，識別性能が高く，また学習に必要な演算量が少ないため学習が高速であるという利点がある．しかし，カーネル関数という間接的な手法により高次元空間への写像を行っているために写像先の高次元空間$\Phi(\ )$の解析が困難であるという欠点を有する．

3. 特　徴　選　択

3・1　特徴選択とは

パターン認識で扱うデータは，高次元である場合が多い．高次元データをそのまま識別器械で識別を行うと，計算コストが高くなる，あるいは，識別に関してあまり意味を持たない不要な特徴（ノイズ）により，最良の識別を行うことができないといった問題が生じる．

n次元のパターンは，n個の特徴量で表されている．このn個の中から単純にm個を選ぶことを特徴選択と言う．最も適当なm個を選ぶことが特徴選択の核心の問題である．特徴選択はパターンがベクトルで表されているとき，パターン空間の次元数を減らす機能を持つ．

特徴選択により識別に関して有効な特徴を取り出すことで，前述の問題を解決することができる．

n個の要素を持つオリジナルの特徴集合を$\boldsymbol{Y}$，所望の選択された部分集合$\boldsymbol{X}$($\boldsymbol{X}\subseteq\boldsymbol{Y}$)の要素数を$d$とする．また，集合$\boldsymbol{X}$に対する評価基準を与える関数を$J(\boldsymbol{X})$と表す．評価値$J$の値が高いほど良い特徴の集合であるといえる．$J(\cdot)$を最大化するので，評価基準関数の1つとして認識率$1-p_e$($p_e$：誤り確率）を用いることができる．ただし評価基準関数として誤り確率を用いることは，特徴選択が用いられる識別器械や，学習データとテストデータのセット数に依存してしまう．

特徴選択問題は，集合$\boldsymbol{X}\subseteq\boldsymbol{Y}$を求めることで，$|\boldsymbol{X}|=d$と，

$$J(\boldsymbol{X})=\max_{\boldsymbol{Z}\subseteq\boldsymbol{Y},|\boldsymbol{Z}|=d}J(\boldsymbol{Z}) \qquad (17)$$

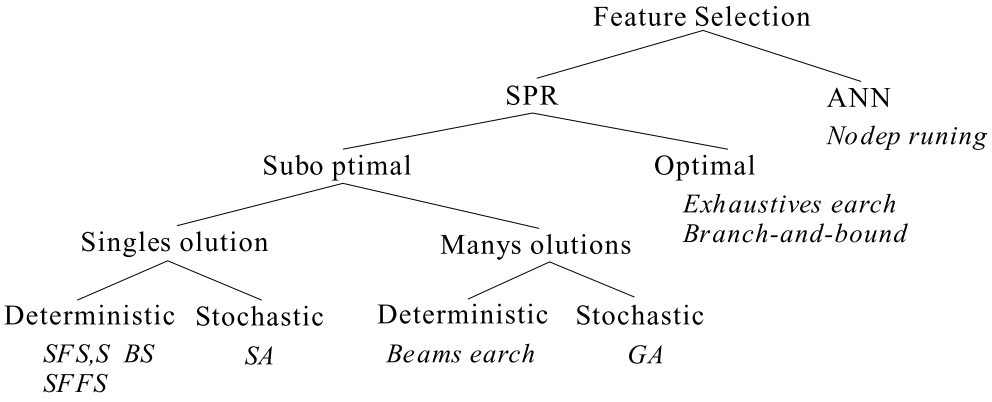

図 1 特徴選択アルゴリズムの分類
Fig. 1 Category of feature selection algorithms.

Input:
$Y = \{y_i \mid j = 1, \cdots, D\}$ //available mesurements//
output:
$X_k = \{x_j \mid j = 1, \cdots, k,\ x_j \in Y\},\ k = 0, 1, \cdots, D$
Initialization:
$X_0 := \emptyset;\ \ k := 0$
Termination:
Stop when k equals the number of features required
Step1 *(Inclusion)*
$x^+ := arg \max_{x \in Y - X_k} J(X_k + x)$
$X_{k+1} := X_k + x^+;\ k := k + 1$
Step2 *(Conditional Exclusion)*
$x^- := arg \max_{x \in X_k} J(X_k - x)$
`if` $J(X_k - x^-) > J(X_{k-1})$ `then`
$X_{k-1} := X_k - x^-;\ k := k - 1$
`goto` **Step2**
`else`
`goto` **Step1**

図 2 SFFS アルゴリズム
Fig. 2 SFFS algorithm.

を満たすものである．

3・2 SFFS（Sequectial Forward Floating Search）

Jain[6]により**図 1** のように特徴選択アルゴリズムが分類された．

その中の “Deterministic, Single-Solution Methods” カテゴリに分けられるアルゴリズムは，単一の解（特徴集合）を持ち，ターゲットとする基準を満たすまで繰り返し特徴を増やしたり減らしたりするものである．これらのアルゴリズムは，特徴選択において最も一般的に使われている手法である．

このカテゴリに属する手法は 2 つのグループに分けることができる．1 つは，特徴数が 0 の状態からスタートし特徴を増やしていく方法であり，forward 型と呼ばれるものである．もう 1 つは，全ての特徴を持つ状態からスタートし特徴を削除していく方法で，backward 型と呼ばれるものである．Foraward 型アルゴリズムの代表例として，Whitney[7]が提案した SFS（sequential forward selection）があり，backward 型アルゴリズムの代表例として，Marill and Green[8]が提案した SBS（sequential backward selection）がある．これらの方法は，可能な全ての特徴の組合せを試せるわけではないので，最適な解を与えるという保証はないということに注意しなければならない．

これらは 1 方向だけの探索アルゴリズムであり，最良の特徴の組合せを求めることは困難である．そこで forward 型と backward 型を組み合わせた floating 型アルゴリズムが研究された．このアルゴリズムの代表例として，Pudil et al.[9]により提案された SFFS（sequential floating forward search）がある．

SFFS のアルゴリズムを**図 2** に示した．SFFS は forward 型をベースとしている．特徴数が 0 の状態からスタートし特徴を増やしていくが，特徴を 1 個増やしたあと，これまでに選択された特徴集合の中から特徴を 1 個削除する．評価値が大きくなる間はこのまま exclusion ステップを続け，評価値が下がれば削除をやめ inclusion ステップに戻る．

4. 実　　験

4・1 実験システム

本研究におけるパターン認識システムは，**図 3** に示すように 2 つのステージにより構成される．第 1 のステージである「特徴選択（Feature Selection）」は，関連のある特徴を選択し，関連のない特徴を省く．第 2 ステージの「識別器械（Classifier）」は，特徴選択により選択された特徴から属するクラスを決定する．

本研究では特徴選択として SFFS を用い，評価基準を SVM による識別率とした．また，識別器械に SVM を用いた．

4・2 実験データ

本研究で用いたデータは，名古屋大学医学部において C 型肝炎患者の血液検査データを収集したものである．それぞれのサンプルは，各患者はあらかじめ肝生検をうけたこ

とで線維化の段階がわかっている．線維化はその度合いによって，F0（正常），F1（軽度線維化），F2（中等度），F3（高度線維化），F4（肝硬変）の 5 段階に分けることができる．実験データはそれぞれの段階の移行段階を含め，F0，F0-1，F1，F1-2，F2，F3，F3-4，F4 と確定診断されており，本研究ではこの診断結果をもとに F0 から F2 をクラス f1，F3 から F4 をクラス f2 とし，これら 2 つのクラス分類を血液検査データから行うことを目的とする．

以上より実験に用いたサンプル数は 217 で，各サンプルに 37 種類の検査項目が含まれている．具体的な検査項目を**表 1** に示す．また，f1 クラスに属するものは 149 パターン，f2 クラスに属するものは 68 パターンである．

4･3　データの正規化

実験を行う前に各特徴量の優位性をそろえるために，データの正規化を行った．

各次元ごとに値が［−1, 1］の範囲となるようにした．本研究で用いたデータにおいて，特徴「HCV 遺伝型」の値は数値データではなく，“1a”，“1b”，“2a”，“2b”，“3c” という値である．各々の値に関して優位性をつけることが困難であるため，各値が同等に扱われるように，**表 2** のように 2 次元の値に変換した．

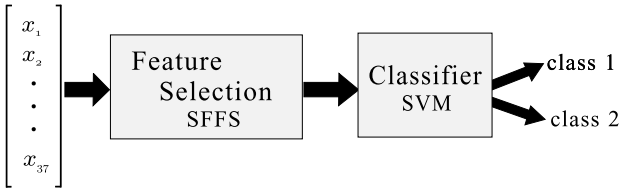

図 3　パターン認識システム
Fig. 3　Pattern recognition system.

4･4　実験方法

特徴選択の SFFS の評価基準に SVM の識別率を用いた．SVM で学習を行うためにはパターン数が多い方が望ましいが，本研究においてはパターン数が 217 であり，十分なサンプル数ではない．

また，全てのデータを学習データとした場合には，識別器械が学習データに特化する overfitting がおきてしまい，十分な特徴選択が行えなくなってしまう．そこで本研究では 1 サンプルをテストデータ，残りのサンプルを学習データとしてテストデータの識別を行う leave one out 法を用いた．これを全てのサンプルがテストデータになるように繰り返すことで識別率を求めた．また SVM の学習には，Vijayakumar［10］による反復的学習法を用いた．反復的学習法のアルゴリズムを**図 4** に示す．

ここで γ は，学習の早さを制御するパラメータである．また，収束の判定は α の変化の割合によって行う．

本研究で学習の終了条件として，α の収束ではなく反復的学習法の Step 2 の繰り返し回数を用いた．

識別能力の評価方法として，本研究で用いたデータはそれぞれのクラスサイズがアンバランスであるため，相乗平均を用いた．全体の識別率（RR_{tot}）は，f1 クラスの識別率を RR_1，f2 クラスの識別率を RR_2 をするとき次式により求める．

$$RR_{\mathrm{tot}} = \sqrt{RR_1 \times RR_2} \tag{18}$$

4･5　実験結果

本実験における SVM の学習で用いたパラーメータ値は，事前の予備実験の結果をもとに σ=0.8，C=100 とした．実験結果を**図 5** に示す．

グラフの縦軸が全体の識別率，横軸が SFFS の繰り返し回数である．グラフより SFFS のループ回数が 234 回のと

表 1　検査項目
Table 1　Observed items.

No.	observed item	No.	observed item	No.	observed item
1	HCV-RNA quantity［k copy/mL］	14	total cholesterol［mg/dL］	27	PT［%］
2	total protein in serum［g/dL］	15	GOT［IU/L］	28	PT time［sec］
3	albumin［%］	16	GPT［IU/L］	29	APTT［%］
4	albumin［g/dL］	17	LDH［IU/L］	30	ferritin［U/mL］
5	α1-globulin	18	ALP［IU/L］	31	TTT［Unit］
6	α1-globulin［g/dL］	19	Γ-GTP［IU/L］	32	ZTT［Unit］
7	α2-globulin［%］	20	LAP［IU/L］	33	hyaluronic acid［ng/mL］
8	α2-globulin［g/dL］	21	CHE［IU/L］	34	sex
9	β-globulin［%］	22	total bilirubin［mg/dL］	35	age［year］
10	β-globulin［g/dL］	23	GOT/GPT	36	infection discovery time［year］
11	γ-globulin［%］	24	platelet［$\times 10^3$/mm^3］	37	HCV genotype
12	γ-globulin［g/dL］	25	serum iron［μg/dL］		
13	A/G	26	HPT［%］		

き最大の識別率 93%を得た．そのときのエラー数は f1 クラスが 11 サンプル（/149 サンプル），f2 クラスが 4 サンプル（/68 サンプル）であった．また，そのとき選択された特徴を**表 3** に示す．

これらの特徴のみで SVM により識別を行うことで 93%の識別率を得ることができた．

このとき，各 SVM において作成されたサポートベクトルの数は平均で 145 個，最大 153 個，最小 134 個であった．各学習には 216 個の学習サンプルがあたえられているため，平均で約 67%のサンプルがサポートベクトルとなった．

表 2　HCV 遺伝型の値変換
Table 2　Value conversion of HCV genotype.

original	relabeled
1a	(1, 1)
1b	(1, −1)
2a	(−1, 1)
2b	(−1, −1)
3a	(0, 0)

1. Initialize $\alpha_i = 0$
Compute matrix $D_{i,j} = y_i y_j (K(\boldsymbol{x_i}, \boldsymbol{x_j}) + \lambda^2)$
for $i, j = 1, \cdots, l$)
2. For each pattern, $i = 1, \cdots, l$, compute
2.1 $E_i = \sum_{j=1}^{l} \alpha_j D_{ij}$
2.2 $\delta\alpha_i = \min\{\max[\gamma(1 - E_i), -\alpha_i], C - \alpha_i\}$
2.3 $\alpha_i = \alpha_i + \delta\alpha_i$
3. If the training has converged, then `stop` else `goto` **Step2**

図 4　反復的学習法
Fig. 4　Sequential algorithm for classification.

Gauss カーネルを用いた NSVM においては，テストデータを各サポートベクトルとの距離という別種の多次元ベクトルに変換して用いるため，作成された識別器械を解析することは困難である．そこで本研究では提案手法と比較するために識別機械として 1-nearest neighbor（以下 1-nn）手法による実験を行った．

1-nn 手法とはテストデータと各学習データの距離を計算し，もっとも距離の小さい学習データの属するクラスを識別結果とする手法である．このため 1-nn は演算に必要なメモリ量が多いという問題点はあるものの，学習データに関しては識別率 100%が保証されている手法であり，識別器械の評価を行う場合の比較対象として一般的に用いられている．

この 1-nn 手法により全特徴を用いた場合と，SFFS で最適とされた特徴のみを用いて識別した場合，それぞれ識別率は 66%，78%であった．一方，特徴選択を行わず全ての特徴を用いて SVM により識別を行うと 76%の識別率であった．以上の結果を**表 4** にまとめた．

この結果より，SVM と 1-nn を比較した場合，特徴選択のありなしにかかわらず SVM が高い識別率を示している．これは 1-nn が全ての学習データを用いることで非常に複雑かつ厳格な識別関数を形成しているのに対して，SVM はソフトマージン最大化を目的としているために，多少の誤差を許容するなめらかな識別境界を作成するため，未学習データに対する汎化性能が大幅に向上しているものと考えられる．また，SVM，1-nn いずれの場合においても特徴選択をすることで大幅な識別率の向上がはかられており，血液検査データを用いた肝炎の病態識別において本研究で提案する特徴選択，すなわち検査項目の選別手法の有効性が示された．

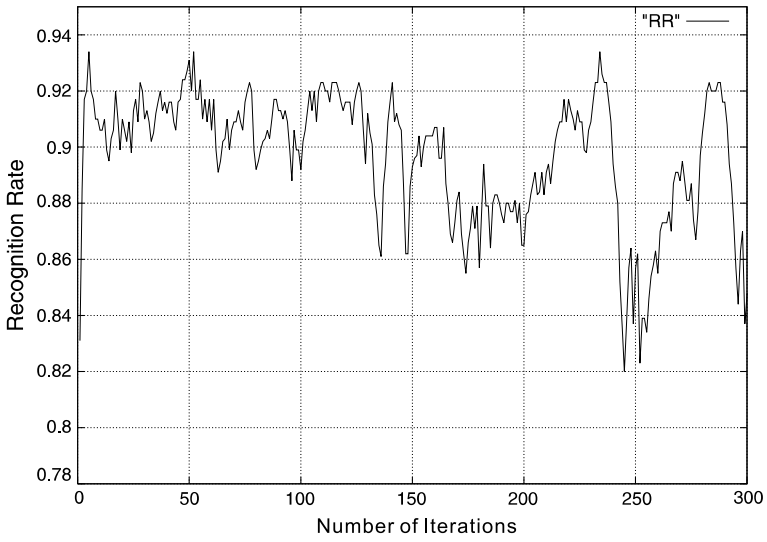

図 5　SFFS による全体の識別率の推移
Fig. 5　Variation of total recognition rate using SFFS.

そして，SVM に特徴選択を組み合わせた場合，未学習のテストデータに対して高い識別率が得られ，本手法が C 型肝炎の病態識別に有効な手法であることが確認された．C 型肝炎の病態 F0 ～ F4 は本来連続的に変化するものであり，これら病態間に絶対的な区別があるわけではない．しかし，本実験は肝生検などを用いて医師によって確定診断が行われた結果をもとにしており，この確定診断結果という離散的な値を最適解として用いている．よって本実験結果に関しても妥当な数値と考えることができる．

さらに実験結果を確認するため，**図 6** に提案手法における F0，F1，F2，F3，F4 の各クラスに属するデータの出力値 f の分布をヒストグラムとして示した．この図より各クラスのピークが左から順番に並び，F0 ～ F2（f1）は負の値を F3 ～ F4（f2）が正の値を出力している様子が確認できる．SVM の出力値 f の絶対値は識別境界（$f=0$）からの距離に比例する．よってこの結果より，提案手法によって形成された識別器械が単に各サンプルを f1，f2 の 2 クラスに分類するだけではなく，線維化の程度に応じて出力値が変化していることが確認できた．

SVM のような識別器械を使うと，次元数を削減することをしなくても overfitting を避けることができると言われている[11]．しかし Guyon[12]によっても言われているように，特徴選択を行うことによって SVM を識別器械とした場合でも，識別率を向上させることができるということが本実験によって分かった．

本実験で生成されたサポートベクトルの数は平均で 145 個である．これは 1-nn で必要なベクトルの数（すなわち学習データの数）の 65%であり，今後学習データが増えた場合にはサポートベクトルの数が診断時の演算コストの増加につながるおそれがある．これについては山口らによりサポートベクトルの数を近似的に大幅に減らす手法[13]が提案されており，この手法を用いることで解決が可能であると考えられる．本手法におけるサポートベクトルの削減に

第 3 表　特徴選択により選択された特徴
Table 3　Selected features using feature selection.

No.	observed item
1	HCV-RNA quantity［k copy/mL］
6	α1-globulin［g/dL］
7	α2-globulin［%］
8	α2-globulin［g/dL］
9	β-globulin［%］
11	γ-globulin［%］
13	A/G
14	total cholesterol［mg/dL］
16	GPT［IU/L］
18	ALP［IU/L］
19	Γ-GTP［IU/L］
20	LAP［IU/L］
22	total bilirubin［mg/dL］
24	platelet［$\times 10^3$/mm^3］
26	HPT［%］
27	PT［%］
28	PT time［sec］
29	APTT［%］
31	TTT［Unit］
33	hyaluronic acid［ng/mL］

第 4 表　実験結果のまとめ
Table 4　Result of examination.

	with SFFS	no SFFS
SVM	93%	76%
1-nn	78%	66%

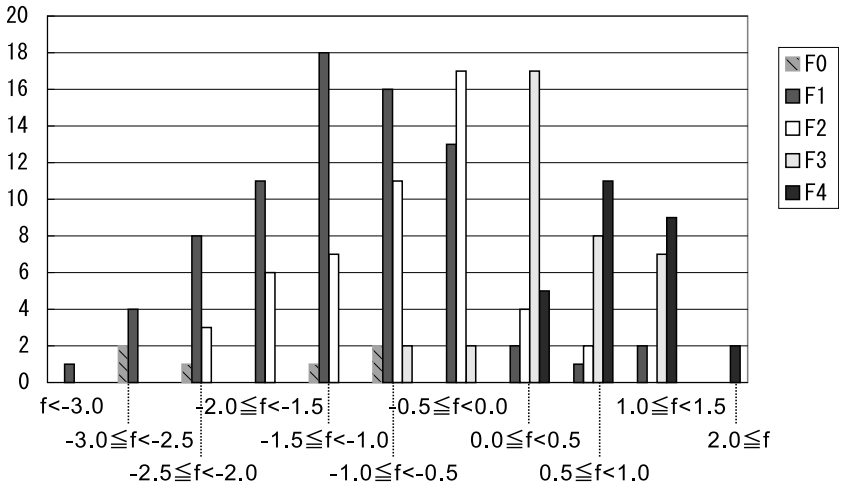

図 6　SVM の識別関数 f 値の線維化段階による変化
Fig. 6　The histograms of the SVM's output.

ついては今後の研究課題としたい．

5. ま　と　め

本研究では統計的識別手法を用いてＣ型肝炎患者の血液検査データから病態を識別するためのシステムを提案した．従来手法の医師による推定での識別精度の低さや，確定診断を得るために行う肝生検での手間や危険性が問題であった．

本研究では，特徴選択による検査項目の選別に，パターン認識の中で最も優れているといわれる SVM を組み合わせることで病態を判別する手法を提案した．SVM を単独で識別器械として用いたときは 76%の識別率であったが，SFFS を組み合わせることにより，識別に不要な特徴を削除することで 93%の識別率を得ることができた．

本実験で用いた SVM はカーネル関数を用いて間接的に高次元空間への写像を実現しているため，得られた識別関数の医学的意味などの解析が困難であるという問題点がある．今後はこの SVM において得られた識別関数の解析手法を検討していきたい．

謝辞　本研究で用いた血液検査データの取得・整理に関して，名古屋大学大学院医学系研究科医療管理情報学の江征先生，ならびに藤田保健衛生大学病院消化器内科吉岡健太郎先生に多大なる御協力をいただきました．ここに感謝いたします．

文　　献

1. Desmet VJ, Gerber M, Hoofnagle JH, Manns M, Scheuer PJ: Classification of chronic hepatitis: Diagnosis, grading and staging. Hepatology. **19**: 1513–1520, 1994.
2. Icida F, Tsujii T, Omata M, Ichida T, Inoue K, Kamimura T, Yamada G, Hino K, Yokosuka O, Suzuki H: Classification report: New Inuyama classification for histological assessment of chronic hepatitis. Int Hepatol Commun. **6**: 112–119, 1996.
3. Batts KP, Ludwig J: Chronic hepatitis: An update on terminology and reporting. Am J Surg Pathol. **19**(12): 1409–1417, 1995.
4. 小栗宏次, 岩田　彰, 深津俊明, 山内一信, ニューラルネットによる慢性肝疾患診断支援システム「NISYS」. 医用電子と生体工学. **32**(2): 106–111, 1994.
5. Cortes C, Vapnik V: Support vector networks. Mach Learn. **20**: 273–297, 1995.
6. Jain A, Zongker D: Feature selection: Evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell. **19**(2): 153–158, 1997.
7. Whitney AW: A direct method of nonparametric measurement selection. IEEE Trans Comput. **20**: 1100–1103, 1971.
8. Marill T, Green DM: On the effectiveness of receptors in recognition system, IEEE Trans Inf Theory. **9**: 11–17, 1963.
9. Pudil P, Novovičová J, Kittler J: Floating search methods in feature selection. Pattern Recognit Lett. **15**(11): 1119–1125, 1994.
10. Vijayakumar S, Wu S: Sequential support vector classifiers and regression. Proc of Int Conf Soft Computing, 1999, pp. 610–619.
11. Vapnik VN: Statistical Learning Theory. Wiley Interscience, Hoboken, New Jersey, 1998.
12. Guyon I: Gene selection for cancer classification using support vector machines. Mach Learn. **46**: 389–422, 2002.
13. 山口拓真, 丸山　稔: 階層型識別器を用いた情景画像からの文字抽出手法. 電子情報通信学会論文誌. **J88DII**(6): 1047–1055, 2005.

青木　一真（アオキ　カズマ）
平成 15 年名古屋工業大学電気情報工学科卒業．平成 17 年同大学大学院工学研究科情報工学専攻博士前期課程了．同年中部電力（株）入社．
日本生体医工学会会員．

黒柳　奨（クロヤナギ　ススム）
平成 3 年名古屋工業大学電気情報工学科卒業．平成 5 年同大学大学院博士前期課程，平成 8 年同博士後期課程修了．博士（工学）．同年名古屋工業大学電気情報工学科助手．平成 15 年同大学大学院助手，現在に至る．パターン認識，聴覚情報処理に関する研究に従事．
日本生体医工学会，電子情報通信学会，日本音響学会，日本神経回路学会会員．

岩田　彰（イワタ　アキラ）
昭和 48 年名古屋大学工学部電気工学科卒業．昭和 50 年同大学大学院修士課程修了．同年名古屋工業大学情報工学科助手．講師，助教授を経て，平成 5 年同大学電気情報工学科教授．平成 14 年同大学副学長，平成 16 同大学大学院教授，現在に至る．工学博士．ニューラルネットワーク，情報セキュリティに関する研究に従事．電子情報通信学会論文賞（1993 年），情報処理学会 Best Author 賞（1998 年）など受賞．
日本生体医工学会，電子情報通信学会，日本神経回路学会など各会員．IEEE Senior Member.

山内　一信（ヤマウチ　カズノブ）
名古屋大学医学部を 1969（昭和 44）年卒業．1974 ～ 1976 年米国ミネソタ大学留学．1979 年から同大学医学部附属病院カルテ部助手，1991 年から同病院医療情報部教授，2000 年から名古屋大学大学院医学系研究科医療管理情報学教授．専門研究分野は電子カルテの開発，情報処理特に意思決定支援システムの開発．
所属学会は日本生体医工学会，日本医療情報学会，日本病院管理学会，日本診療録管理学会などである．