

サポートベクターマシンを用いた肺腫瘍の鑑別

山田 菜美*, 橋 理恵**, 木戸 尚治***

Classification of Pulmonary Nodules Using Support Vector Machine

Nami YAMADA, Rie TACHIBANA and Syoji KIDO

Abstract

Accurate segmentation of small pulmonary nodules (SPNs) is an important technique for volumetric doubling time estimation and for feature characterization of the diagnosis of SPNs. Therefore, we have developed an automated volumetric segmentation algorithm of SPNs on thoracic CT images. This paper presents quantitative evaluation of the shape of the segmented SPNs on thoracic CT images, and also presents quantitative estimation of the growth rate of SPNs by use of repeated thoracic CT images. To quantitatively evaluate the performance of our algorithm, we employed the coincident rate. This coincident rate was calculated with both the computerized segmented region of an SPN and the region traced by a chest radiologist on axial, coronal MPR, sagittal MPR images. As the result of 90 selected cases, the mean value of the total coincident rate was 0.75 ± 0.081 . We also estimated the performance of our algorithm by use of the doubling time of SPNs on repeated CT images. The doubling time of SPNs obtained from computerized segmented regions closely correlated with the image findings reported by radiologists. From these results, we consider that our algorithm is useful for quantitative estimation for the shape and the growth rate of SPNs on repeated CT in daily clinical practice.

Key words: small pulmonary nodule, segmentation, repeated CT, doubling time

1. はじめに

現在、日本における喫煙率は増税や taspo の導入により減少傾向にあるが、諸外国と比べると、未だ高い状況にある。また、厚生労働省の報告によると、肺がんによる死亡者数は 1 年間で 6 万人を超え、年々増加傾向にある。そこで早期発見、早期治療はがんによる死亡者数の抑制に効果があると考えられており、最近では、マルチスライス CT 装置が人体を広範囲にわたって 1mm 以下の解像度で撮影することが可能となり、小さな腫瘍などの早期発見も可能となっている。しかし、それに伴い患者 1 人あたりの CT 画像が増加している。加えて、CT 画像での診断では読影医により診断結果が異なる事や見落としを防ぐために複数の医師による二重読影が理想と

されていることから読影医は多大な労力と時間を強いられている。この疲労による見落としが懸念されていることから、計算機を第 2 の意見として使用する計算機支援診断システムへの期待が寄せられている。

このような社会的背景をふまえて、医用画像の計算機支援診断システムの開発が盛んに行われており [1]、孤立性肺腫瘍の良悪性鑑別に関する研究も多く行われている [2-4]。胸部 X 線 CT 画像における肺腫瘍の所見では、腫瘍のいびつさや辺縁や内部の濃淡分布などが鑑別の指標となる。先行研究ではこのいびつさ、複雑さといった人間の主観量を定量化し、

その値を用いた解析により肺腫瘍の良性・悪性や異常・正常のような 2 クラスへの分類を試みた報告がある。しかし、先行研究[4]では特徴量に CT 値の平均や分散といった統計量や腫瘍の体積など多くの特徴を用いているにも関わらず識別率が十分に高いとは言えない。また、第 2 に意見として計算機を使用する事を目指す場合、医師が診断するときと同じように、他クラスへの分類が必要であると考えられる。

そこで本稿では特徴量に統計量ではなく腫瘍内部の濃淡ヒストグラムや度数分布を特徴に用いることで、各クラスの特徴をより顕著に表現できるのではないかと考えた。この特徴を用いて胸部 X 線 CT 画像上の腫瘍に対して良悪性鑑別が可能であるかに加え、疾患の種類を指摘することができればより質の高い診断が期待できることから良悪性の 2 クラスだけではなく多クラスへの分類を行う。このとき決定木を用いることでより高い識別率を得る事ができるかを検討する。

なお、今回使用した腫瘍データは、CT 画像特徴に基づきあらかじめ読影医によって良性腫瘍が充実型(solid)と充実型以外(nonsolid)の 2 クラス、悪性腫瘍が充実型(solid)、混合型(partsolid)、スリガラス型(nonsolid)の 3 クラスに分類されたものである。また識別器を選択するにあたって、一般に、特徴ベクトルの次元数を D とすると、各クラスの学習データは $4D \sim 5D$ 個必要であると言われているのに対し、今回十分な数の学習データを用意する事が出来なかった。

そこで学習データが少なくても高い識別率を得る事ができるサポートベクターマシン(Support Vector Machine,SVM)[5]を用いた。SVM は 2 クラスの分類問題を解くために作られた教師あり学習機械であるが、複数組み合わせることによって多クラスへの分類を行うことが可能となる。

2. 対象と方法

2.1 対象データ

実験に用いた画像データは、小型肺腫瘍を対象として撮影された CT 画像である。以下に詳細を記す。図 1～図 5 は腫瘍を含む関心領域画像例であり、図の下に説明文は、各グループの画像濃淡特徴である。

今回実験に用いたデータはすべて確定診断がなされたものであり、データ数は 146 症例で良性腫瘍が 60 例(solid : 31 症例, nonsolid : 29 症例)、悪性腫瘍が 86 例(nonsolid : 37 症例, partsolid : 28 症例, solid : 21 症例)を含み、 partsolid とは nonsolid と solid の両方の性質をもつ症例とした。シリーズ枚数は 1 症例あたり 36～47 枚であった。

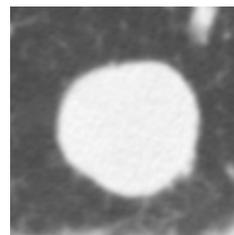


図1. 良性 solid
(腫瘍全体に高濃度 CT をもち、石灰化を有する場合がある)

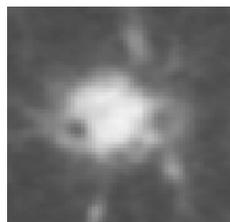


図2. 良性 nonsolid
(腫瘍の内部に高濃度 CT を持ち、辺縁部にスリガラス陰影をもつ)

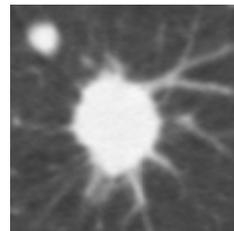


図3. 悪性 solid
(腫瘍全体に高濃度の CT 値を持つ)

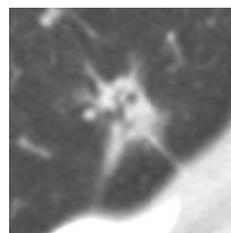


図4. 悪性 partsolid
(腫瘍の内部に高濃度の CT 値を持ち辺縁部にスリガラス陰影を持つ)

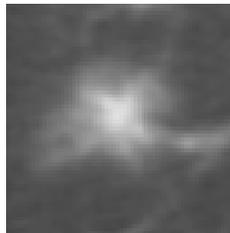


図5. 悪性 nonsolid
(腫瘍全体にスリガラス陰影が広がっている)

また、各画像サイズは 512*512pixel, 画素サイズは 0.325mm, CT スライス厚は 2mm, スライス間隔は 1mm であった。

2.2 処理手順

大まかな処理の流れを次に示す。本稿では、特徴量として CT 画像上の腫瘍内部の濃淡ヒストグラムの相対度数分布を用いた。しかし図 6 のように CT 画像には、腫瘍以外に血管や胸壁などの周辺構造物が含まれている。そのためまず複数の CT 画像を読み込み、関心領域の設定を行い、血管などの周辺構造物を除去する領域抽出を行い、肺腫瘍領域を得る。この抽出された肺腫瘍領域から特徴量の計算を行い、識別を行った。次項より処理について詳しく説明していく。

2.2.1 関心領域の設定と領域抽出

関心領域とは、腫瘍が撮影された陰影の領域など注意を払って観察することが望まれる領域をいう。本研究で使用した領域抽出法[6]は関心領域画像に対して適用するアルゴリズムであるため、GUI を用いて医師により正確に関心領域を設定されたものを用いた。

図 6 の左図が原画像であり、右図が腫瘍を対象として切り出された関心領域画像である。領域抽出を適用し、図 7 のように血管などの周辺構造物を削除した腫瘍のみの画像を得た。

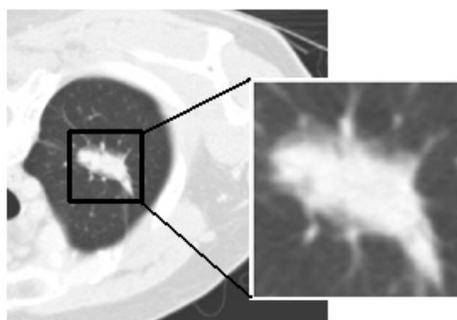
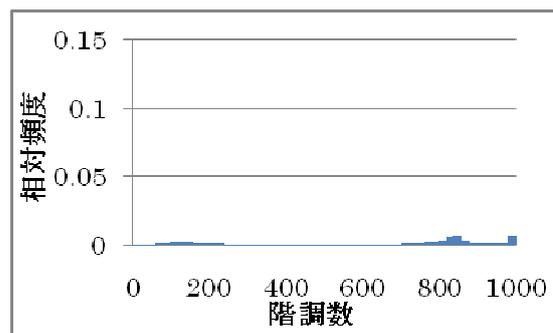


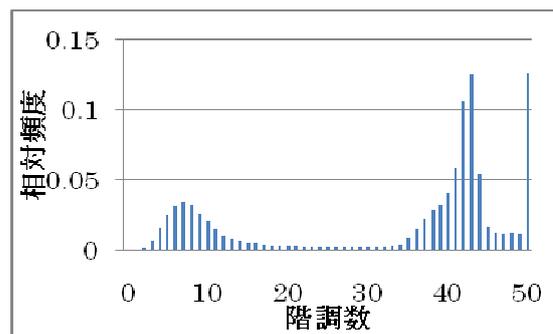
図6. 原画像からの関心領域の設定(左図に原画像, 右図は関心領域画像)



図7. 肺腫瘍の領域抽出



(a)



(b)

図8. 原データのヒストグラム(a)と階調値変換後のヒストグラム(b)

2.2.2 特徴量の計算

図 7 のように肺腫瘍の領域抽出を行った画像から特徴量を計算した。濃淡ヒストグラムは画像全体の各濃度値に対して同じ濃度値を持つ画素数を求め、度数化したものであり画像の濃度値に関する情報をとらえることができる。

本研究で用いた CT 画像の各画素値は CT 値 (Hounsfield unit;HU)で示されており、この人体の大部分を占める水の X 線吸収係数を 0 として規格化したものである。CT 値は水を基準の 0H.U.として設定されている。そのため腫瘍は約 -800H.U.~200H.U.の CT 値で示されており、そのまま約 1000 階調で表現した場合、図 8 の(a)に示すように、次元数が非常に高くなり特徴がつかみにくい。そこで、図 8 の(b)に示すようにヒストグラムに対して階調値変換を行い、原画像 1000 階調ヒストグラムの 20 階調区間にある度数を 1 階調区間に変換することで 50 階調に低減した。このヒストグラムの各濃度区間あたりの画素の度数を特徴値とした。これにより、次項で述べる識別器への入力ベクトルは $\mathbf{x} = (x_1, x_2, \dots, x_i, \dots, x_{50})$ となる。このとき x_i は i 番目の濃度区間に含まれる画素の相対度数である。本研究で使用した各クラスの 2 次元画像の例と濃淡ヒストグラムの例を図 9~図 13 に示す。このとき、縦軸が相対頻度を表し、横軸は階調数を表している。

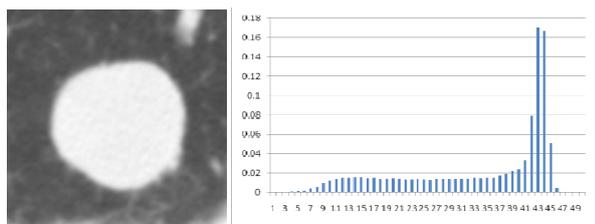


図9. 良性 solid の 2次元画像(左)と濃淡ヒストグラムの度数分布(右)

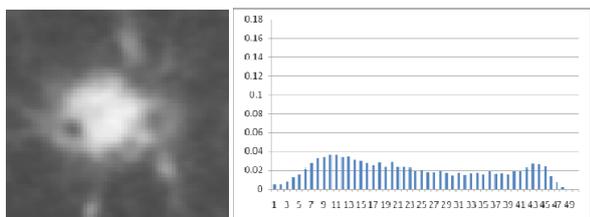


図10. 良性 nonsolid の 2次元画像(左)と濃淡ヒストグラムの度数分布(右)

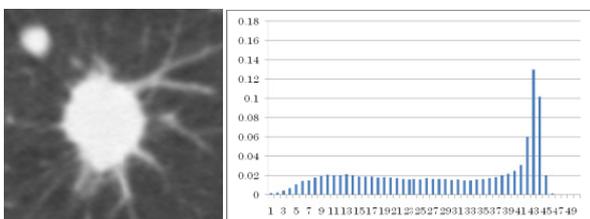


図11. 悪性 solid の 2次元画像(左)と濃淡ヒストグラムの度数分布(右)

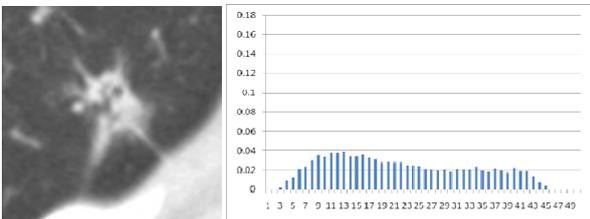


図12. 悪性 partsolid の 2次元画像(左)と濃淡ヒストグラムの度数分布(右)

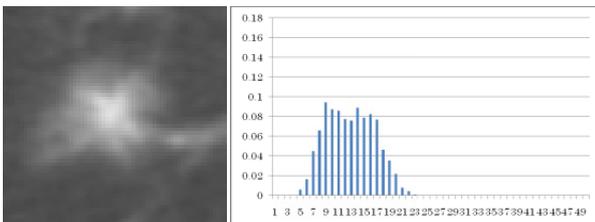


図13. 悪性 nonsolid の 2次元画像(左)と濃淡ヒストグラムの度数分布(右)

なお、左図では 2次元画像を示しているが、右図のヒストグラムは腫瘍の 3次元画像から算出したものであるため、左図からは読み取ることのできない濃淡値も含まれている。

各グループのヒストグラムが持つ特徴について述べる。図 9～図 13 に示す左図の濃淡を比べると、良性 solid と悪性 solid が似ているが、右図のヒストグラムを比較すると、良性 solid には石灰化と呼ばれ

る病状が含まれるため高濃度値が他に比べて高い。悪性 solid も似たヒストグラムを持つが、悪性 solid は石炭化がないため、良性 solid より低濃度値を多く持っている。同じように左図より良性 nonsolid と悪性 partsolid と悪性 nonsolid がの 3 グループが似ている。しかし、右図のヒストグラムでは良性 nonsolid と悪性 partsolid は低濃度値も高濃度値も持っていてよく似ているが、悪性 nonsolid については低濃度値が非常に高い点の特徴である。

2.2.3 識別方法

クラス分類問題を考える際は、多くの識別方法の中からどれを選択するかが問題となるが、本研究では十分な数の学習データを用意する事が出来なかったため、学習データが少なくても高い識別率を得る事ができる SVM を識別器として用いた。

SVM はクラスとクラスの領域を区切る分離超平面からサンプルまでの距離(マージン)が最大化するように分離超平面を構築しクラス分類を行う方法であり、高い汎化能力を持つ。また、SVM は後述の 1), 2) で定義されるように、線形の識別器であるが、3) に示すようにカーネル関数を組み合わせることによって非線形に拡張でき、SVM を複数組み合わせることで 4) に示すように多クラスの分類が可能である。

1) 線形 SVM

まず線形の SVM について述べる[7]。今、 M 個の m 次元の学習データ $x_i (i = 1, \dots, M)$ があると、クラス A 又は B に属するとする。クラス A なら教師信号 y_i を $y_i = 1$ 、クラス B なら $y_i = -1$ とする。この学習データを m 次元入力空間から写像関数 $g(x)$ により l 次元空間に写像した空間で線形分離可能な識別関数 $D(x)$ は次式で与えられる。

$$D(x) = w^T g(x_i) + b, \quad i = 1, \dots, M \quad (1)$$

ここで w は l 次元重みベクトル、 b はバイアス項と呼ばれる定数である。このとき、学習データを完全に識別する超平面は無限に存在する。そこで、超平面を一意に定めるために、

$$w^T g(x_i) + b \begin{cases} \geq 1, & y_i = 1 \\ \leq -1, & y_i = -1 \end{cases} \quad (2)$$

の制約条件をもうける。この式は

$$y_i (w^T g(x_i) + b) \geq 1, \quad i = 1, \dots, M \quad (3)$$

と等価になり、このとき超平面

$$D(x) = w^T g(x) + b = c \quad -1 < c < 1 \quad (4)$$

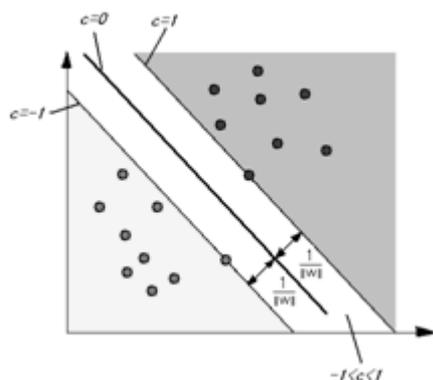


図14. 二次元空間での最適分離超平面

は2クラスの分離超平面を構成し $D(\mathbf{x}) = 0$ は $-1 < c < 1$ で最大マージンをもつ超平面となり、これを最適分離超平面と呼ぶ。また、境界上に存在するデータをサポートベクトルと呼ぶ。図14に式(3)を満たす2つの超平面を示す。

つまり、線形SVMの問題は図14に示すように $1/\|\mathbf{w}\|$ を最大化する問題となり、最適分離超平面を決めるための目的関数が次式で与えられる最適化問題に帰着する。

$$\min Q(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 \quad (5)$$

$$\text{s.t. } y_i(\mathbf{w}^T g(\mathbf{x}_i) + b) \geq 1, \quad i = 1, \dots, M \quad (6)$$

2) ソフトマージン

しかしながら、実際にはほとんどの分類問題が特徴空間において線形分離可能ではない。そこで解決する手法としてソフトマージンや非線形SVMの2つが挙げられる。

ソフトマージンとは、線形分離不可能な場合に許容解を存在させることで最適化を行う。ソフトマージンは \mathbf{x}_i に対して非負のスラック変数 $\xi_i (\geq 0)$ を導入し、式(3)を次式のように変更する。

$$y_i(\mathbf{w}^T g(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad i = 1, \dots, M \quad (7)$$

これにより解くべき最適化問題は以下ようになる。

$$\min Q(\mathbf{w}, b, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + r \sum_{i=1}^M \xi_i \quad (8)$$

ここで、 $\xi = (\xi_1, \dots, \xi_M)^T$ で $r (> 0)$ はマージンパラメータである。式(8)の第1項はマージンを最大化する項として働き、第2項は正となるスラック変数の和あるいは二乗和になるため誤識別したデータが少なくなるように作用し、マージンパラメータは両者の

目的関数のトレードオフを決定する。このため最大の汎化能力を実現するためには、パラメータの値の設定が重要であるが、汎用的にパラメータを設定する手法は現在確立していない。しかし、パラメータの値は特徴ベクトル数より小さくなくてはならないと言われている。

そこで本稿では使用した特徴量が50であったことから、パラメータ r の値を40に設定した。

3) 非線形SVM

次に非線形SVMについて述べる。SVMの非線形識別はカーネルトリックによって入力空間を高次元空間に写像することにより、線形分離可能な状態となり、識別面を求めることが可能となる。このときの識別関数 $D(\mathbf{x})$ は

$$D(\mathbf{x}) = \sum_{i=1}^M \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \quad (9)$$

となる。 $K(\mathbf{x}_i, \mathbf{x})$ はカーネル関数であり、 α_i はラグランジュ乗数である。

なおカーネル関数にはいくつかの種類があるが、本研究では次式のRBFカーネルを用いた。

$$K(\mathbf{x}_i, \mathbf{x}) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}\|^2) \quad (10)$$

その理由として多項式カーネルやシグモイド関数と違い、パラメータによって有効でない場合が存在しないことや調整するパラメータが1つですむため容易であり、パラメータ γ をうまく調整すれば、カーネル関数を用いない場合と同程度の計算量ですむことがあげられる。

4) 多クラスSVM

最後に多クラスSVMについて述べる[8]。本研究では良悪性の2クラスではなく良性を2クラス、悪性を3クラスの計5クラスの識別を行う。そのため2クラスSVMを拡張した多クラスSVMが必要である。

多クラスSVMには大きく分けて、1対多SVMとペアワイズSVMの2つが存在する。これらは n クラス問題を解くにあたって、1対多SVMでは各クラスに対し、1つのクラスと残りのクラスというように、残りのクラスを1クラスと考え、これらの分離超平面を求める事で多クラスへの分類を行う。そのため n 個の決定関数を決めることになる。ペアワイズSVMではクラス1とクラス2というように2クラスずつの分離超平面を求める。このとき $3 \sim n$ の残りのクラスは作用しない。これをすべてのクラスのペアの分離超平面を求める事で分類を行う。そのため $n(n-1)/2$ 個の決定関数を決める必要がある。しかしながら、クラス対に対応する決定関数を決める

際は、先ほど述べたように 2 クラスのデータのみを使って学習するため 1 対多 SVM では分離超平面を決定する際にすべての学習データを使用しなければならないのに対し、ペアワイズでは一度の計算に使用する学習データ数が少なくすむため、短い学習時間での計算が可能となる。なお、本稿ではペアワイズ法により複数クラスの分離超平面の決定を行うライブラリ libSVM[9]を使用した。

3. 実験と結果

3.1 実験

はじめに、今回特徴量として用いた濃淡ヒストグラムの度数分布が良悪性の識別にどの程度有用であるかを調べるために、先行研究と同じように各良性腫瘍と各悪性腫瘍とのペアを 2 クラス SVM での識別実験を行った。

次に、第 2 章の 2.1 節で示したデータを用いて複数クラスを分類する多クラス SVM での識別実験を行った。このとき分類クラスとして良性 solid, 良性 nonsolid, 悪性 solid, 悪性 partsolid, 悪性 nonsolid の 5 つのクラスを用いた。

まず分類方法として、図 15 に示すように一度に 5 クラスへの分類と、図 16 と図 17 に示すように多段階で分類する方法について検討する。多段階で分類する方法として、本研究で使用した 5 つのクラスは病理分類から見ると大きく良性・悪性に、濃淡特徴からみると solid・nonsolid・partsolid に分けられる。そこで第一段階として良性か悪性かで分類し、第二段階として濃淡特徴により識別する決定木(図 16)と、反対に第一段階で濃淡値の高い solid と濃淡値の低い solid 以外に分類した後に、第二段階で良性と悪性に識別を行う決定木(図 17)で検討を行った。

3.2 結果

先行研究の結果と実験により得た 2 クラス分類の誤識別率を表 1 と表 2 に示し、多クラス SVM を用いて複数クラスへ分類した際の識別率を表 3～表 5 に示す。表 1 は良性 solid と各悪性腫瘍、表 2 は良性 nonsolid と各悪性腫瘍のペアを分類した際の誤識別率である。表 3 は一度に 5 クラスへの分類を行った際の各クラスの識別率と全体の識別率を表している。表 4 は良悪性で分類した後に各クラスへ分類する決定木 2(図 16)、表 5 は濃淡特徴で分類した後に各クラスへ分類する決定木 3(図 17)に従って分類実験を行った時の識別率である。第一段階の結果のように分かれているセルは左のセルが各クラスの識別率で、右側が決定木においてのその節全体の識別率を示す。以下はその識別率を求める式である。こ

の式で示すように、第一段階及び第二段階に示されている識別率は各段階のみの識別率であり、最終識別率は第一段階、第二段階を経ての識別率である。

$$\text{第一段階} = \left(\frac{\text{第一段階で正しく識別されたデータ数}}{\text{テストデータ数}} \right)$$

$$\text{第二段階} = \left(\frac{\text{第二段階で正しく識別されたデータ数}}{\text{第一段階で正しく識別されたデータ数}} \right)$$

$$\text{最終識別率} = \left(\frac{\text{第一・二段階で正しく識別されたデータ数}}{\text{テストデータ数}} \right)$$

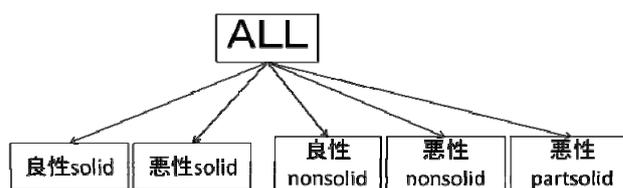


図15. 決定木 1 (5 クラスの分類)

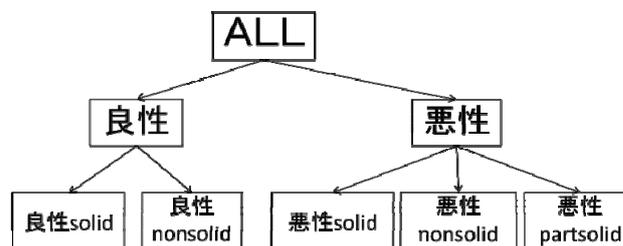


図16. 決定木 2 (良悪性後、各クラスへの分類)

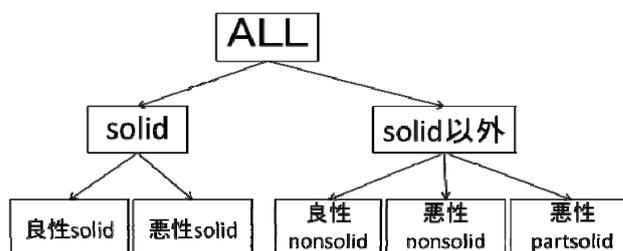


図17. 決定木 3 (濃淡特徴で分類後、各クラスへの分類)

表1. 先行研究と本研究の誤識別率
(良性 solid と各悪性腫瘍の 2 クラス分類)

クラス名	先行研究の誤識別率%	本研究の誤識別率%
悪性 solid	28.0	20.0
悪性 nonsolid	5.1	3.4
悪性 partsolid	10.9	15.6

表2. 先行研究と本研究の誤識別率
(良性 nonsolid と各悪性腫瘍の 2 クラス分類)

クラス名	先行研究の誤識別率%	本研究の誤識別率%
悪性 solid	25.0	7.7
悪性 nonsolid	19.2	14.8
悪性 partsolid	43.9	42.4

4. 考察

はじめに、表 1 と表 2 に示す各良性腫瘍と各悪性腫瘍のペアを分類した結果では先行研究と比べると 6 パターンのうち 5 パターンが誤識別率を下げる事が出来た。また先行研究では濃淡平均や分散、コントラストや形状特徴などたくさんの特徴を用いていたのに対し、本稿では腫瘍内部の濃淡ヒストグラムの度数分布のみを特徴に用いた。このことから、先行研究では腫瘍の濃淡を平均や分散などの統計量を用いていたが、ヒストグラムの度数分布値を特徴に用いた方が各クラスの特徴をより顕著に示せたと考ええる。

次に多クラス SVM を用いて複数クラスへの分類を行った実験結果について考察する。まず決定木 1(図 15)のように、一度に 5 クラスの分類を行った識別結果(表 3)より、最も低い識別率は悪性 solid で 52.4%、最も高い識別率は良性 solid で 86.2%とばらつきが見られ、全体の識別率は 67.1%と低かった。この原因として、各クラスのヒストグラムを比べると、クラスが違っていても似たヒストグラムをもつものも含まれたため、識別面が上手く生成できなかったことが挙げられる。

それに加え、一度に 5 クラスへの分類では他のクラスに誤って分類されてしまう確率も高い。これらのことから 5 クラスを一度に識別する事は難しかったと考えられる。また、第 2 の意見として計算機支援診断システムとして使用する場合、5 年生存率の低い悪性 solid の識別率が低いことは致命的な問題であることから決定木 1 の方法は良悪性鑑別に適切でないと考ええる。

つづいて決定木 2(図 16)や決定木 3(図 17)のように多クラスの分類を段階的に行った際の識別結果(表 2, 表 3)について考察する。すべてのデータを大きく 2 つのクラスに分類した第一段階では、良性と

表3. 決定木 1(図 15)のときの識別率

クラス名	各クラス(%)	全体(%)
良性 solid	86.2	67.1
良性 nonsolid	54.8	
悪性 solid	52.4	
悪性 partsolid	62.9	
悪性 nonsolid	76.7	

表4. 決定木 2(図 16)のときの識別率

	クラス名	第一段階 (%)		第二段階(%)		最終識別率 (%)	
良 性	良性 solid	71.7		100.0	86.8	86.2	71.9
	良性 nonsolid			83.3		48.4	
悪 性	悪性 solid	90.7	82.9	93.8	86.8	71.4	71.9
	悪性 nonsolid			80.0		80.0	
	悪性 partsolid			91.4		74.3	

表5. 決定木 3(図 17)のときの識別率

	クラス名	第一段階 (%)		第二段階(%)		最終識別率 (%)	
solid	良性 solid	74.0		95.8	81.3	79.3	71.2
	悪性 solid			100.0		61.9	
solid 以外	良性 nonsolid	94.8	87.7	70.4	81.3	61.3	71.2
	悪性 nonsolid			80.0		80.0	
	悪性 part-solid			73.5		71.4	

悪性に分類した識別率は全体で 82.9%、濃淡特徴から solid と solid 以外へ分類した識別率は全体で 87.7%となり、第一段階においては濃淡特徴に違いを持つ solid と solid 以外へと分類した決定木 3 の方が識別率の高い結果となった。一方、第二段階では、第一段階の結果とは反対に第一段階で良悪性により

分類を行った決定木 2 の識別率の方が 86.8%と高かった。この結果より、いずれの段階においても、濃淡特徴の分類を行った決定木で高い識別率を得る事ができた。この要因として、本稿では特徴に濃淡情報のみを用いたことが挙げられる。そのため、良悪性分類を行う段階では良性和悪性で異なる特徴を持つ形状などを特徴量として使用し、濃淡特徴による分類を行う段階では濃淡特徴量を使用する事でより高い識別率を得ることができるのではないかと考える。

最終識別率を比較すると段階的に分類を行った決定木 2 と決定木 3 において大きな差は見られなかった。また、第一段階で濃淡特徴に違いを持つ solid と solid 以外で分類した際、悪性 solid のほとんどが solid 以外と誤識別されてしまったため、最終識別率は 61.9%と非常に低かった。同様に、第一段階で良悪性による分類を行った決定木 2 では良性 nonsolid の最終識別率が 48.4%と非常に低かった。このように、第一段階で良性 solid と同じクラスに分類されたクラスは最終識別率が非常に低い結果となった。このことから、識別結果には良性 solid の特徴が大きく影響していると考えられる。実際に画像を見ると、良性 solid は石灰化を見とめる場合が多い。つまり、濃度値の高い値を持つことになり、この特徴が識別面の生成に大きく影響したと考える。本稿では 5 クラスを分類するにあたって、一度に分類するのではなく、決定木による多段階的に分類する方法について検討した。その結果、段階的に分類を行った方が各クラスの識別率は高くなり、良悪性で分類した後に各クラスへの分類を行う決定木 2 において悪性 solid の識別率が最も高い結果となった。また、本稿では第一段階で 2 対 3 への分類を行っている。しかし実験結果より、良性 solid のように強い特徴を持つグループから分類を行っていく 1 対 4 の分類を行うことで良い結果を得る事が出来るのではないかと考える。また、濃淡特徴だけでは、上手く分類する事が出来なかった。実際の診断では辺縁の特徴も重視されていることを考えても、今後は形状の特徴量を含めた識別を検討していく必要があると考える。

5. まとめ

本研究では、特徴量を腫瘍内部の濃淡ヒストグラムの度数分布とし、5 クラスの肺腫瘍を分類し良悪性鑑別を行った。また識別器として SVM を用いたが 5 クラスの分類は難しいと考え、決定木を用いることで識別問題を分けて考え、最適な構造モデルの検討を行った。結果より充実性で分類した後に各クラスへの識別を行う手法が有用であった。

今後の課題として、より最適な決定木の検討を行うことやヒストグラムに対して濃度階調値変換を行う際に、今回では 50 階調としたが、最適な階調数を求める事や、形状特徴などを取り入れることでより識別率の高い特徴ベクトルの検討を行うことや、最適なパラメータの設定を行うことが課題となる。

そして今回の実験では、決定木の節ごとに識別実験を行い、その結果から次の節の実験を行ったため、その作業の自動化などが挙げられる。

参考文献

- [1] 飯沼武: 癌検診におけるコンピュータ支援診断 (CAD) の役割. Med Imag Tech, Vol. 20, No6, pp671-675, 2002
- [2] 平野靖, 他: 3 次元胸部 CT 像からのインタラクティブな肺腫瘍領域抽出と良悪性鑑別への応用. 信学論. J87-D-II, No1, pp237-247, 2004
- [3] 木戸尚治, 他: CT における肺腫瘍性病変のフラクタル解析 良性腫瘍と悪性腫瘍の鑑別. 臨床放射線, Vol. 43, No1, 1998
- [4] 一宮博幸: SVM を用いた肺腫瘍の鑑別. 平成 15 年度山口大学工学部卒業論文, 2004
- [5] 麻生英樹: パターン認識と学習の統計学. 岩波書店, 2003
- [6] 橋理恵, 他: 胸部 X 線 CT 画像を用いた小型肺腫瘍の領域抽出. 信学論. J87-D-II, No1: pp228-236, 2004
- [7] 阿倍重夫: パターン認識のためのサポートベクトルマシン入門-I-2 クラス SVM. システム/制御/情報, Vol. 52, No7: pp245-250, 2008
- [8] 阿倍重夫: パターン認識のためのサポートベクトルマシン入門-II-多クラス SVM. システム/制御/情報, Vol. 52, No9: pp340-345, 2008
- [9] LIBSVM, <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>