3B1-R-2-4

SVMと学習データ選択によるクラス分類アルゴリズムの検討

Examination of the class separation algorithm of the data based on SVM and learning data selection

大堀 裕一[*1] 廣安 知之[*2] 横内 久猛[*2]
Yuichi OBORI Tomoyuki HIROYASU Hisatake YOKOUCHI

[*1]同志社大学大学院生命医科学研究科
Graduate School of Life and Medical Sciences, Doshisha University

[*2]同志社大学生命医科学部
Department of Life and Medical Sciences, Doshisha University

In this research, the data of breast cancer is classified. In the conventional way, a malignant growth and benign tumor of data are classified. On the other hand, in this research, the data which is definitely belonging to malignant growth area and the data which has the possibility to be belonging to benign tumor area are divided. Because the libraries of conventional two value classification learning method have high ability and easy to use, the proposed system uses two value classification learning is utilized to achieve the goal. This problem is formulated as combinatorial optimization problem. In this formulation, the malignant growth area is maximized. In this paper, the algorithm of this problem is described and the effectiveness of the proposed method was discussed through the numerical experiments.

1. はじめに

癌の診断において，患者の腫瘍が悪性であるか良性であるかという判断は病理診断によって行われる．病理診断とは，生体から摘出した組織を薄く切りだして染色し，顕微鏡で観察することによって病変の有無や種類を診断する方法である．病理診断は，病理学の知識や医師の経験によって診断を行う専門性の高い技術である．そのため，診断する医師によって結果が異なるといったことや，医師の経験の違いによって精度が異なるといった問題点が存在する．現在，この病理診断によって得られた過去の患者の腫瘍のデータが蓄積されている．そこで過去の患者の腫瘍のデータを使用して，医師がある患者の腫瘍が良性か悪性か診断する際に補助となる情報を提示するツールが求められている．腫瘍のデータとは，ある患者の腫瘍が良性または悪性であるかを示し，その腫瘍の病理画像から得られたいくつかの特徴量を持つものである．しかし，悪性腫瘍と良性腫瘍のデータの分布が重複する領域がある場合，良性と悪性の領域に分けても両方の領域において誤分類が存在し，患者の腫瘍データが入力された際良性，悪性どちらであるか正確に判断できないという問題が存在する．そこで本研究では，悪性，良性という2つのクラスに完全に分離することが困難なデータに対して，過去のデータから良性または悪性であると確実に判断できる領域とどちらか判断できない領域に分ける事を提案する．すると，正確に判断できなかったデータに対しても，一部領域においては確実に判断することが可能となる．

本稿では，従来の 2 値分類学習法ではクラスの完全な分離が困難であり，未知データのクラスの正確な判断ができない場合においても，一部領域において判断可能とする提案手法について述べた後，提案手法を用いた実験とその結果を示す．

連絡先: 大堀 裕一，同志社大学大学院 生命医科学研究科 医情報学専攻，京都府京田辺市多田羅都谷 1-3，0774-65-6924，yohbori@mis.doshisha.ac.jp

2. 良性悪性のクラス分類

2.1 病理画像の特徴量を用いた分類

本研究では，人体から摘出した腫瘍の病理画像から得られたデータを使用している．このデータは，ある腫瘍が良性であるか，悪性であるかというクラスと，その腫瘍の病理画像から抽出したいくつかの特徴量を値として持つ．ある病理画像から得られたデータは特徴空間と呼ばれる d 次元空間の 1 点として表現できる．この d 次元ベクトルを特徴ベクトルと呼ぶ．上記のデータを良性と悪性の 2 クラスに分類する問題として，入力と出力間の関数を，与えられたデータから学習する方法を考える．学習とは学習データを利用して，未知の特徴ベクトルがどちらのクラスに属するか判定する関数を求めることである．本研究では，患者の腫瘍病理画像の特徴量を入力とし，その腫瘍が良性であるか悪性であるかということを出力とする関数を学習することを考える．学習は特徴空間上で，2 クラスの識別線を決定することに相当する．この考え方は，パターン認識手法と呼ばれる．

2.2 SVM

本研究では，識別能力に優れている SVM(Support Vector Machine) をパターン認識手法として用いる．SVM の識別線から最も近いデータをサポートベクトル，識別線からサポートベクトルまでの距離をマージンという．SVM における学習とは，図 1 のようにマージンが最大となるような線形識別線を計算する事に相当する．SVM は汎化能力に優れており，カーネルトリックによって非線形に拡張することが可能である [1][2][3].

3. 提案手法

3.1 提案手法の概要

提案手法では，結果を視覚化するため 2 次元データで行う．図 2 のような，良性のデータと悪性のデータが重複する部分が存在する学習データの場合，完全なクラスの分離が困難であり，従来の方法で良性と悪性のクラスに分けたとしても誤った分類が存在してしまい，診断の精度としては信頼性が低下す

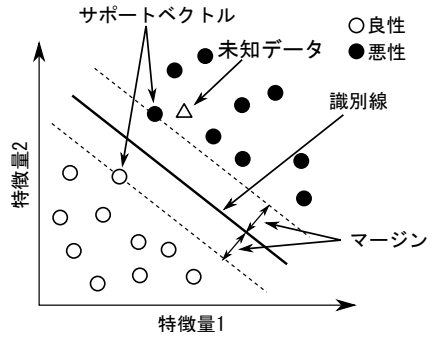

図 1: 2 次元データにおける SVM

る．そこで，悪性と良性のデータが重複した領域は，クラスの判断が不可能な領域として，クラスが重複していない判断可能領域とクラスが重複している判断不可能領域に分ける手法を提案する．これによって，一部領域において良性または悪性であるか過去のデータからは正確に判断が可能となる．さらに，判断可能領域を広くすることが求められる．これは，提案した領域に分離できるように学習データの組み合わせを選択することで実現する．例えば，悪性と判断できる領域と判断不可能な領域に分離したい場合，悪性の学習データの組み合わせを選択する．図 3 は悪性のデータを判断可能領域として選択した例である．するとクラスの分離が不可能であったデータが，図 3 のように過去のデータから悪性と判断できる領域と，どちらか判断できない領域に分離することができる．同様に，良性と判断できる領域と，どちらか判断できない領域に分離することも可能である．学習データ選択の方法としてはデータの組み合わせ最適化問題と捉え，最適化アルゴリズムによって行った．最適化アルゴリズムには，組み合わせ最適化問題に適した遺伝的アルゴリズム (Genetic Algorithm:GA)[4][5] を用いた．GA は自然界における生物の進化のメカニズムを工学的にモデル化した最適化アルゴリズムであり，問題に適した解を効率よく探索することができる．

また，例のように 2 クラスを持つデータには多数の特徴量が存在する．良性または悪性であることにあまり関係していない特徴量で提案手法を行うより，関係性の高いもので行った方が有効であると考えたため，データの中で良性または悪性であることに最も関係している特徴量を 2 つ選択し，提案手法を行った．

3.2 提案アルゴリズム

以下に提案するアルゴリズムの流れを説明する．

Step 1 2 次元データとして視覚化

データを視覚化するため，特徴量を 2 つとし 2 次元空間として分布を示す．

Step 2 特徴量選択

良性，悪性であることに最も関わっていると考えられる特徴量を 2 つ選択する．選択された特徴量で示されるデータに対して学習データ選択を行う．

Step 3 学習データ選択

図 3 のように学習するデータを選択することで，良性または悪性のデータのみ存在する判断可能領域と，良性と悪性のデータが混在する判断不可能領域に分離できる．さらに，判断可能領域はできる限り広くする．以上の条件を満たす学習データの組み合わせを求めるため，まず多数の学習データの組み合わせを生成し，GA によって問題に適したデータの組み合わせの探索を行う．

Step 4 SVM による識別線表示

SVM によって判断可能領域と判断不可能領域に分離する識別線を表示する．

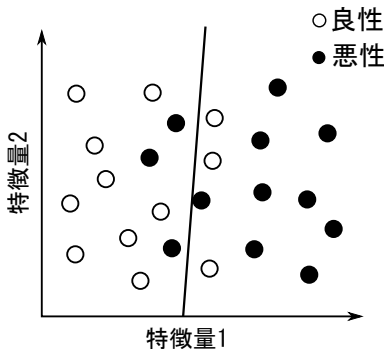

図 2: 学習データの例

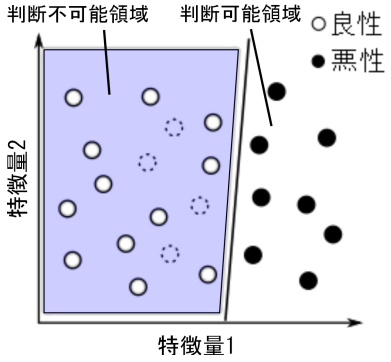

図 3: 提案手法の概要

3.3 提案手法の実現

3.3.1 特徴パラメータの選択

良性と悪性のデータの分布が離れている特徴量であるほど，その特徴量は良性まはた悪性であることに関わっていると考えらえる．そこでパラメータの中から 2 つを選び，その全ての組み合わせにおいて，それぞれのクラスの重心を求め，クラスの重心間のマハラノビス距離 [6] が最大となったパラメータの組み合わせを選択した．腫瘍のデータは特徴量によってスケールが異なるため，ユークリッド距離は無効となるため，特徴量のスケールの違いと，良性と悪性のそれぞれのクラスの分散を考慮したマハラノビス距離を使用した．マハラノビス距離の式を以下に示す．

$$d = \sqrt{(G_a - G_b)^T S^{-1} (G_a - G_b)} \tag{1}$$

ただし，G_a，G_b はクラスの重心ベクトル，S^{-1} は共分散行列の逆行列である．

3.3.2 最適化問題の定式化

n 個のデータが与えられているとする．設計変数を学習データの組み合わせとして，選択するデータを 1，選択しないデータを 0 でデータ長の 2 値ビット配列で表し，図 6 のように学習データに対応させる．n 個の中から k 個が選択されたとき，

学習データ数は k 個となる．

与えられた学習データのクラスと，SVM による分類が異なれば誤識別となる．学習データ数が k 個のとき，学習データに対する誤識別率を Err とする．判断可能領域の広さを学習データの数で表現すると，学習データ数が大きくなるほど領域は広くなる．そこで，選択された学習データは全て正確に識別するという制約条件で，選択される学習データ数 k を最大化する．制約条件を満たさない場合，ペナルティとして減じる．

$$max \quad k \tag{2}$$

$$subject\ \ to \quad Err = 0 \tag{3}$$

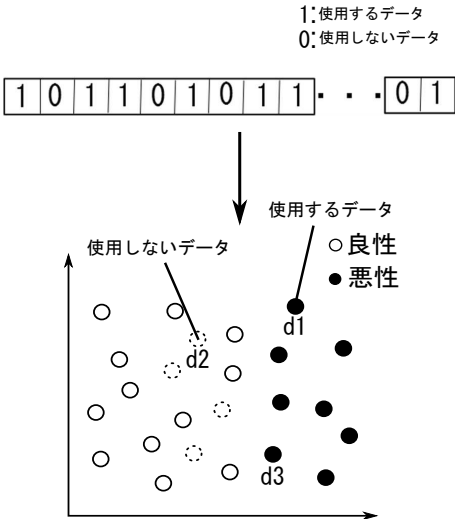

図 4: データ選択の例

4. 実験

4.1 実験概要

今回の実験の目的は，良性と悪性のクラスに最も関係していると考えられる特徴量を 2 つ選択し，クラスの完全な分離が不可能なデータに対して，提案手法を用いて悪性と判断できる領域と，どちらか判断不可能な領域に分けられるかを確認することである．以下にアルゴリズムの流れを説明する．まず，使用する腫瘍のデータの特徴量の中から 2 つを選び，2 次元空間として表現する．この操作を特徴量全ての組み合わせにおいて行う．次に，全ての場合において良性のデータと悪性のデータの重心をそれぞれ求め，式 (1) に示したマハラノビス距離を算出する．マハラノビス距離が最大となった特徴量をクラスとの関係性の高いものとして用いる．そこで，SVM によって悪性の領域と判断不可能領域に分離が可能となる学習データの組み合わせを GA によって探索し，確認を行う．GA で用いたパラメータを Table.2 に示す．また，SVM による識別線は 2 次元多項式とした．

4.2 実験データ

データセットは，UCI Machine Learning Repository の breast cancer[7] のデータを用いた．このデータは乳房組織から腫瘍を採取し，その腫瘍が良性であるか悪性であるかというクラスを示し，腫瘍の病理画像から抽出した 10 個の特徴量を有している．596 個のデータの中から 400 個を対象データとした．breast cancer データの特徴量を Table.1 に示す．

表 1: GA のパラメータ

パラメータ	値
世代数 [個]	150
個体数 [個]	200
選択手法	トーナメント選択
トーナメントサイズ	4
交叉方法	一様交叉
交叉率 [rate]	0.9
突然変異率 [rate]	0.01

表 2: 腫瘍データの特徴量

	特徴量
1	radius
2	texture
3	perimeter
4	area
5	smoothness
6	compactness
7	concavity
8	concave points
9	symmetry
10	fractal dimension

4.3 実験結果

4.3.1 特徴量選択の結果

Table.1 に示した全てのパラメータの組み合わせにおいてマハラノビス距離を求めた結果，texture と concave points の場合に最大値 6.39 を示した．texture は細胞画像の 1 ピクセルのグレースケール値の標準偏差を示し，concave points は細胞核の凹みの割合を示している．また，symmetry と fractal dimension の場合に最小値 0.55，全組み合わせにおいての平均値は 3.71，標準偏差は 2.66 であった．図 7 にマハラノビス距離が最大値となった特徴量を用いたデータの分布，図 8 にマハラノビス距離が最小値となった特徴量を用いたデータの分布を示す．

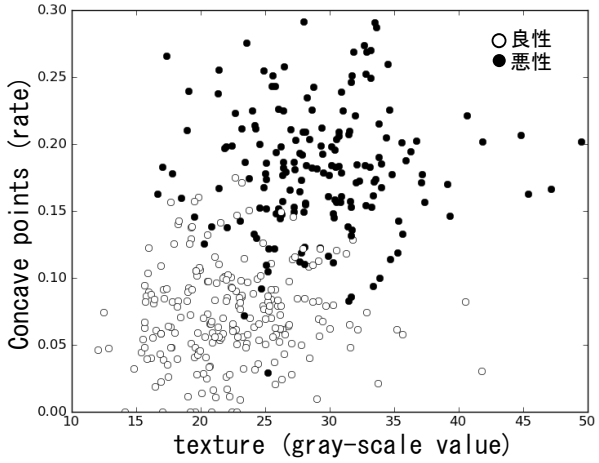

図 5: 選択された特徴量による腫瘍データ

4.4 提案手法による実験結果

図 9 に従来の方法で実験を行った結果，図 10 に提案手法を用いた実験結果を示す．従来の方法で良性のクラスと悪性のクラスに分けたとき，良性の領域において誤識別数が 17 個，悪性領域においても誤識別数が 17 個となった．提案手法を用いた場合，悪性の領域において誤識別数は 0 となった．

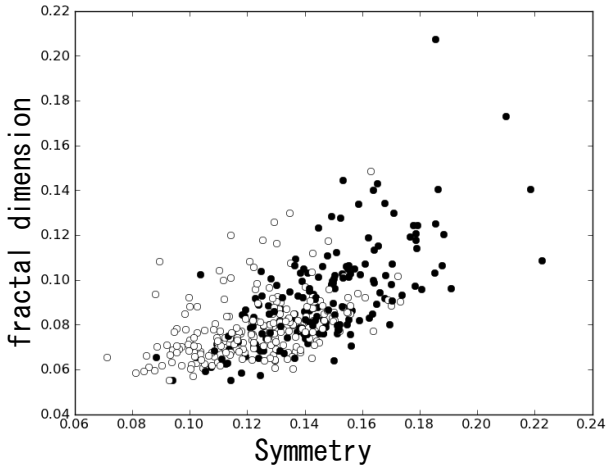

図 6: クラスとの関係性が低い特徴量による腫瘍データ

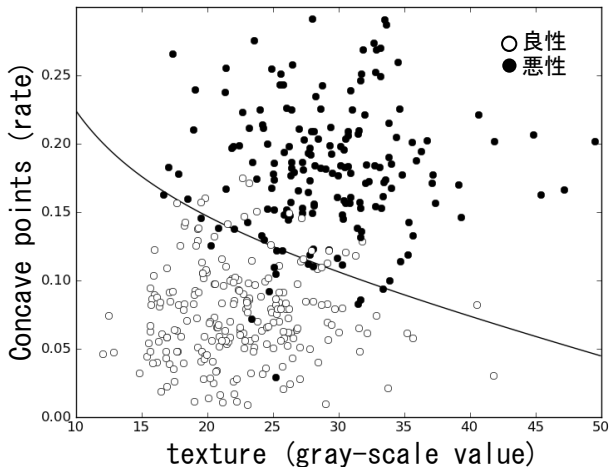

図 7: 従来の手法の結果

5. 考察

特徴量選択の結果において，マハラノビス距離が最大値となった特徴量を用いたときのデータの分布図と最小値となった特徴量を用いたときのデータの分布図を比較すると，前者の図の方が悪性と良性のデータの分布が離れていることが分かる．このことから，良性と悪性であることに関わっている特徴量を選択できたと考えられる．また，従来の方法で良性と悪性の領域に分けた場合，良性と悪性の領域両方において誤った分類が存在するため，新たに診断する患者の腫瘍データが入力されるとき，その腫瘍が良性であるか悪性であるか正確な判断が困難となる．提案手法を用いた場合は，悪性の領域において誤識別が 0 となった．従って，患者の腫瘍のデータがこの領域に入力された場合，過去のデータからは悪性であると判断できる．

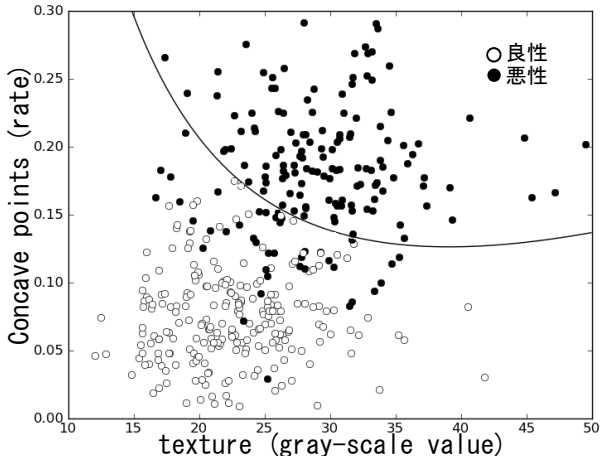

図 8: 提案手法の結果

以上から，悪性と判断できる領域とどちらであるか判断不可能な領域に分離することができたと考えられる．

6. まとめ

本稿では，医師が患者の腫瘍の診断を行う際の補助となる情報を提示するため，患者の腫瘍の病理画像から得られたデータを学習して，新たに診断を行う患者の腫瘍が良性であるか，悪性であるか判断する方法を提案した．しかし，良性と悪性の 2 つのクラスに完全に分離することが困難な場合，SVM でこの 2 つのクラスに分けたとしても信頼性が低下する．そこでこういったデータに対して，一方のクラスのデータを選択し，どちらのクラスであるか過去のデータから判断が可能な領域と，判断不可能な領域に分ける手法を提案した．この問題はデータの組み合わせ最適化問題と捉え，学習データの組み合わせを設計変数とし，誤識別が 0 という制約条件で学習データ数を最大化する最適化問題とすることで実現した．その際，良性・悪性であるということに最も関係していると考えられる特徴量を選択した．この上で，breast cancer の病理画像のデータを用いて実験を行った結果，判断可能領域と判断不可能領域に分けることができた．

参考文献

[1] 阿久津達也,”バイオインフォマティクスの数理とアルゴリズム”, 共立出版,2007

[2] 津田宏治,”サポートベクターマシンとは何か”, 電子情報通信学会誌,Vol.83,No6,pp.460-466,2000

[3] Nello Cristianini,John Shawe-Taylor,”サポートベクターマシン入門”, 共立出版,2005

[4] 棟朝雅春,”遺伝的アルゴリズム：その理論と先端的手法”, 森北出版,2008

[5] 伊庭斉志,”遺伝的アルゴリズムの基礎-GA の謎を解く”, オーム社,1994

[6] 杉山一成, 奥村学,”半教師有りクラスタリングを用いた Web 検索結果における人名の曖昧性解消”, 言語処理学会論文誌 自然言語処理, Vol.16, No.5, pp.23-49, 2009

[7] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html