東北薬科大学研究誌，56，67－74（2009）
Journal of Tohoku Pharmaceutical University，56，67－74（2009）

医療データ解析へのサポートベクトルマシン（SVM）の応用

青木　空眞，佐藤　憲一，星　　憲司，川上　準子，森　　弘毅，[a] 齋藤　芳彦，[b] 吉田　克己 [a]

[a] 東北大学大学院医学系研究科，[b] 東北大学病院

Applying Support Vector Machine in the Analysis of Medical Data

Sorama Aoki, Kenichi Sato, Kenji Hoshi, Junko Kawakami, Kouki Mori,[a]
Yoshihiko Saito,[b] and Katsumi Yoshida[a]

(Received November 20, 2009)

In our previous papers we proposed a novel screening method that assists the diagnosis of Graves' hyperthyroidism via two types of neural networks by making use of routine test data. This method can be applied by non-specialists during physical check-ups at a low cost and is expected to lead to rapid referrals for examination and treatment by thyroid specialists, that is, to improve patients' QOL.

In this report, we apply the support vector machine, which is a novel learning method building on kernels, to the classification problems of medical data such as Wisconsin breast cancer data or our screening of hyperthyroid. It turned out that the support vector machine, after best tuning of parameters based on the grid-search method, works quite well to correctly classify the samples located in the bordering area between two classes. Our results suggest that the SVM would work as a useful method in our screening in addition to previous two types of neural networks.

Key words —— screening; routine test data; support vector machine; grid-search method; Bayesian regularized neural network

健康診断や病院初診時にはコレステロールをはじめとする基本的検査が行われるが，甲状腺の検査は基本的検査の項目に含まれていないため，甲状腺機能異常はしばしば見逃されているのが現状であり，中には1年以上も誤診されている場合もある．もし甲状腺機能異常が疑われればホルモン検査などにより確定診断が可能で，治療もしやすい疾患であるが，ホルモン検査のコストが高価なことなどもあり疑わしいということでなければすぐに検査をするのは難しい．そこで，基本的検査項目のセットからパターン認識の手法を用いて甲状腺機能異常のスクリーニングを行い，確実な甲状腺専門検査へとつなげる早期発見のための診断支援ができれば患者 QOL 向上の面でも非常に有用である．

最近，我々はベイズ正則型ニューラルネットワーク（BRNN）[1,2] と自己組織化型ニューラルネットワーク（SONN）[3] の2種類を併用して，バセドウ病患者データと対照の健常者データの基本的検査項目のセットから両者を分類する可能性を検討したところ，アルカリフォスファターゼ（ALP），血清クレアチニン（S-Cr），総コレステロール（T-Cho）の3項目のセットに注目することで甲状腺機能亢進症患者を高い確率で予測することが可能である．[4,5] 現在では，甲状腺機能低下症についても4項目の基本的検査セットに注目することで，亢進症と同様に高い確率で予測することが可能である．[6,7] この手軽な新しいスクリーニング手法が広く病院や健診施設で利用されるようになれば，甲状腺機能異常の患者を早期にスクリーニングして，専門医による確定診断と治療につなげ，患者 QOL の向上を実現できるものと期待できる．

そこで，2008年7月から JR 仙台病院健康管理センターの人間ドック受診者を対象に本格的にスクリーニングを開始しており（JR 仙台病院倫理委員会の承認済），その結果，スクリーニングにより甲状腺機能異常が疑われたため送付した勧奨文により，甲状腺外来を受診された方の中から，バセドウ病患者，甲状腺機能低下症患者，無痛性甲状腺炎，慢性甲状腺炎の方が見つかっており，治療がなされている．

人間ドックでは症状の軽い方も対象となりやすく，2群の分離境界周辺での分別精度のさらなる向上が望まれる．この点で，Vapnik により導入されたサポートベクトルマシン [8] は，“マージン最大化”という優れた原理に基づいており，2群の識別能力が

BRNN を超える可能性も期待できるので，本論文でその可能性を検討してみる．その準備として，まず，正弦関数で分離された2つの領域に属するデータの分類問題，次に，Wisconsin Breast Cancer データの分類問題について，グリッドサーチ法によりサポートベクトルマシンの最適パラメータを決定して，[9] サポートベクトルマシンによる分類の精度を BRNN と比較してみる．最後に，甲状腺機能異常症のスクリーニングに，BRNN およびサポートベクトルマシンを適用した場合の結果を詳しく比較する．

方　　法

サポートベクトルマシン（SVM）

SVM は従来の非線形モデルと比べて，局所解に陥らないことや，マージン最大化に基づく分離超平面により得られる汎化性の高さ，分離超平面近傍に存在するサポートベクトルと呼ばれる一部の学習データしか分類に関与しないことによる計算の速さが特徴である．

SVM では入力データに対して非線形変換を行い高次元空間において分離平面を求めるが，このままでは計算量が膨大となってしまうため，実際にはカーネルトリックと呼ばれる方法により，高次元空間上での計算を避けてカーネル関数を用いた計算に置き換える．代表的なカーネルとしては線形カーネル，多項式カーネル，Radial Basis Function（RBF）カーネル，シグモイドカーネルなどが存在する．

結局，SVM の方法をデータ分類に適用するには，分離境界をまたいでしまう学習データに対して，逸脱をどこまで許容するかを決定する識別エラーに対するペナルティのパラメータ C と，採用する各カーネルに現れる固有のパラメータの両方を，解析するデータに合わせた最適なものに決定するステップが重要である．このパラメータを決定すれば，分類に直接関与する分離境界直近の学習データ（サポートベクトル）が適切に定まり，高速で精度の高いクラス分類が可能となる場合が多いことが知られている．

SVM のフリーソフトとしては，SVM^{light}，TinySVM 等が公開されているが，本研究では Chang らの開発した LIBSVM[10] を用いた．

ベイズ正則型ニューラルネットワーク（BRNN）

BRNN は三層の階層型ニューラルネットワークである．入力層の各ニューロンに検査項目を割り当て，出力層の2つのニューロンは Softmax 型出力をもち，それぞれ，健常者と甲状腺機能亢進症患者の予測確率を与える．既知のサンプルの診断結果を用いて，ネットワークの重みをベイズ学習によって決定するが，マルコフ連鎖モンテカルロ法を用いた実装[3] によりクラス分類能力も高い安定したネットワークであることが知られている．この手法ではベイズの先見的結合荷重確率が結合荷重の正則化項と自然に対応しており，ARD（Automatic Relevance Determination）により判定に大きく影響した一連の入力，すなわち，基本的検査項目を評価することができる．

本研究では Neal の開発した“Software for Flexible Bayesian Modeling and Markov Chain Sampling”package[11] を用いた．

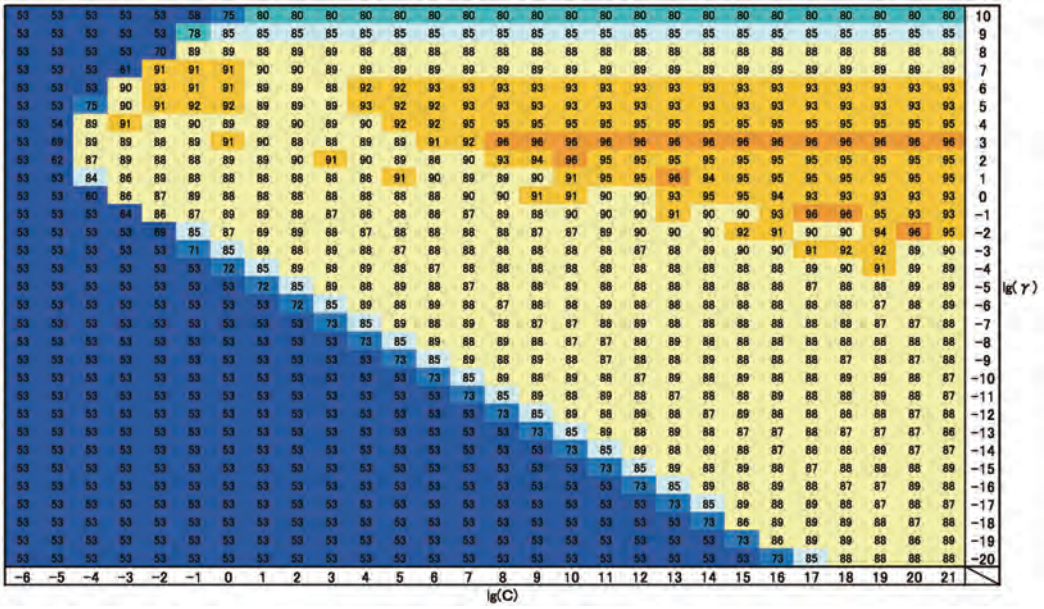
結 果 と 考 察

グリッドサーチによる最適なパラメータの探索

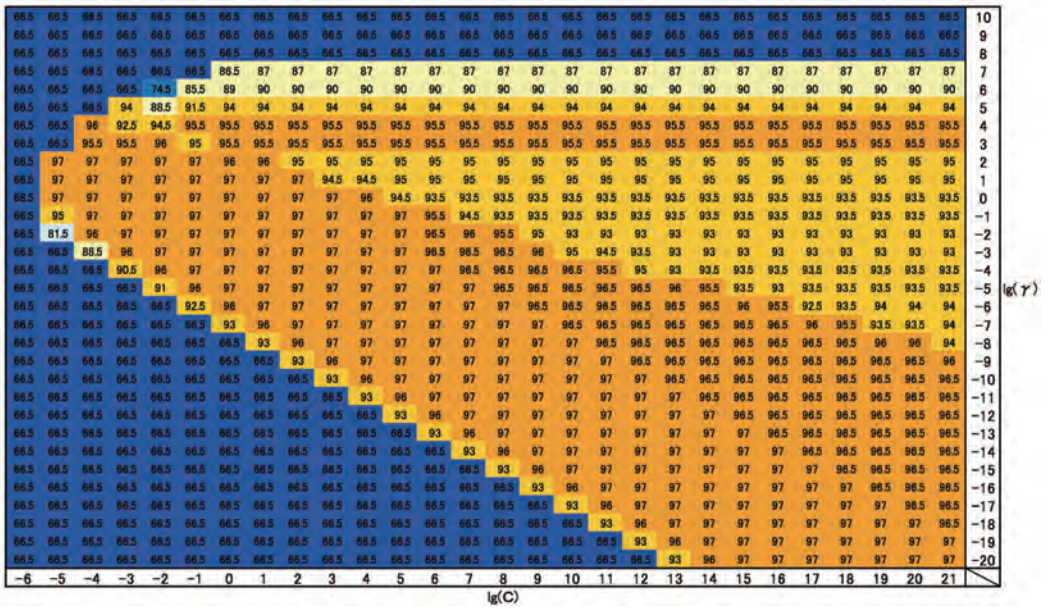
研究ではパラメータの適切な設定が比較的容易であり，ほかのカーネルと比べても同等，あるいはより優れた振舞いをする RBF カーネルを用いた．[12,13] RBF カーネルは n 個の学習データを（x_i, y_i,）,i＝1, 2, …, n（ここで x_i, y_i,は i 番目のサンプルの従属変数 x_i とクラス番号 y_i,〜1 または－1 である）とすると次式で与えられる．

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2)$$

ここで，SVM を使用するにあたり設定しなければならない未定のパラメータはペナルティパラメータ C と RBF カーネルのパラメータ γ の2つである．これらのパラメータは問題に合わせて最適な値に設定する必要があり，今回，我々はこのパラメータを決定するためにグリッドサーチと呼ばれる手法を用いた．これは，パラメータ C と γ を対として，それぞれのパラメータを規則的かつ網羅的に少しずつ動かして計算を行い，結果をプロットしてその中で最も精度の高い組み合わせを選択するやり方である．今回，$C = 2^{-6}, 2^{-5}, 2^{-4}, \ldots, 2^{21}$，$\gamma = 2^{-20}, 2^{-19}, 1^{-18}, \ldots, 2^{10}$ と動かして計算を行い，学習データに対する 10-fold Cross Validation での分類精度をプロットしてパラメータを決定した．グリッドサーチの実例を Fig. 1 に示す．



A. Grid-search result for 100 points of two-dimensional artificial data set is shown. In this data set, the two classes are separated by the nonlinear boundary determined by the Eq. $y = \sin(\pi x)$. 47 points in the upper region are defined as Class 1, while 53 points in the lower region are defined as Class 2.



B. Grid-search result for 200 subjects of Wisconsin Breast Cancer Database. 133 benign cases are defined as Class 1, while 67 malignant cases are defined as Class 2.

Fig. 1. The results obtained by grid-search on tune parameter C $=2^{-6}, 2^{-5}, 2^{-4}, \ldots, 2^{21}$ and $\gamma = 2^{-20}, 2^{-19}, 1^{-18}, \ldots, 2^{10}$

The number labeled on cell in these figures represents the correct rate (%) for training data set, which was calculated by SVM with 10 fold cross validation method.

正弦曲線により2群に分けられた2次元データの分類問題

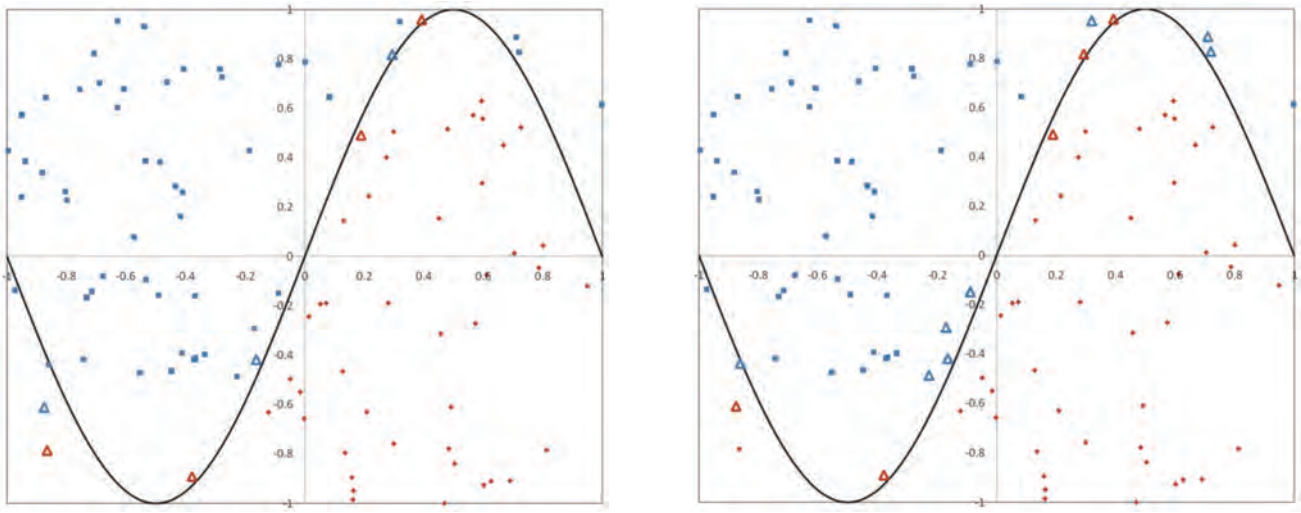
2次元データの非線形分類問題に対して，多変量解析に属する判別分析，BRNN，SVMの3つの手法による分類結果にどのような分類特性の相違が見られるかを検討した．

データとして，x-y平面上-1～1の範囲でランダムに100点をとり，正弦曲線を引いて曲線上部の座標点をクラス1，下部をクラス2と2群に分けて学習データとして，学習を行った．その後，新たな100点のテストデータを作成して予測させ，結果の相違を調べた．Table 1に，3つの手法でのそれぞれの計算条件下における誤りの数を示す．SVMにおいてはサポートベクトル（SV）の数も示した．BRNNでは

Table 1. Predicted results for 100 test points based on the three different models at various parameter values

model		train errors	test errors	#SV
Discriminant Analysis		10	13	−
BRNN (mid12, ARD−)	200 MCMC steps	11	17	−
	400 MCMC steps	1	5	−
	800 MCMC steps	0	5	−
SVM (C−SVC, RBF)	C＝32, $\gamma = 2^{-12}$ Accuracy*: 53%	47	74	94
	C＝1, $\gamma = 1$ Accuracy: 88%	8	12	56
	C＝8192, $\gamma = 2$ Accuracy: 96%	0	5	11

* "Accuracy" shown for SVM means the correct rate (%) for training samples when cross-validation calculation is done using the grid-search.



A. Calculated results with BRNN (the number of neurons in the intermediate layer＝12, without ARD, 800 MCMC samplings).

B. Calculated results with SVM (parameters are set as C＝8192 and $\gamma = 2$, respectively).

Fig. 2. The best result with BRNN is compared to that with SVM

The blue points belong to the class 1, and the red ones to the class 2, respectively. The symbol "△" represents that the point was predicted as the class with the probability of 50-70%.

マルコフ連鎖モンテカルロ法（MCMC）のサンプリング回数が十分でない時は判別分析の結果よりもエラーが多かったが，800 回まで増やすと完全な分類が可能となった．一方，SVM ではグリッドサーチの結果が最も悪いパラメータセットを用いると全データに対してクラス 2 との判定を下してしまうが，最も良好なパラメータセットを用いることで完全な分類が可能となり，SV の数も 11 と少ない優れた解が得られた．

BRNN と SVM での最良な結果同士を詳細に比較してみると，Fig. 2 に示すように，どちらも間違いの少ないクラス分類ができている点では同じだが，BRNN では 7 点を除いて出力値 70%以上で予測，つまりあまり遊びを持たせずに分類しているのに対して，SVM では正弦曲線に沿って緩やかな出力値を持っており，境界についてより適切に取り扱っていることがわかる．

Wisconsin Breast Cancer Database による 9 次元データの２群分類問題

Wisconsin Breast Cancer Database（WBCD）は，米国 Wisconsin 大学病院の Wolberg によって集められた，顕微鏡検査による乳癌患者の細胞データである．[14,15] 1〜10 のランクで評価された Table 2 に示

す9つの項目からなるこのデータには，診断結果として良性（Benign）～クラス1か悪性（Malignant）～クラス2, のどちらであるかが記載されており，クラス番号を教師信号として与えてBRNNとSVMに予測を行わせる．WBCDは検査項目に欠損のあるデータを除くと全部で683名のデータが使用可能であり，この中から200名をランダムに抽出して学習データのグループとした．また，同様にしてさらに200名をランダムに抽出し，こちらをテストデータのグループとした（Table 3）．

Table 2. As for the WBCD data, each subject has 9 attributes which take the integer value between 1 and 10

No	Attributes
1	Clump Thickness
2	Uniformity of Cell Size
3	Uniformity of Cell Shape
4	Marginal Adhesion
5	Single Epithelial Cell Size
6	Bare Nuclei
7	Bland Chromatin
8	Normal Nucleoli
9	Mitoses

SVMのパラメータ選択の手法としてはグリッドサーチを用いるが，正解率をプロットしたのみのグリッドサーチを行った場合では同じ正解率を持つ組み合わせが多く，合理的な選択が困難であった．そこで，このような異常サンプルのスクリーニングにおいては偽陰性の予測を最も避けるべきケースと考え，学習データに対する偽陰性の数をパラメータ毎にプロットしたグリッドサーチも併せて実施し，両者を見て全体の正解率が最も高く，かつ偽陰性の数が最も少ないパラメータを選択することにした（Fig. 3）．

予測の結果をTable 4に示す．BRNNではMCMCサンプリング回数を400回とると結果は既に安定しており，400回と800回の結果を比較すると，誤判定した患者の内訳，予測率はほとんど変化が見られなかった．SVMにおいては，グリッドサーチの正解率は同じ97%であるが偽陰性のグリッドサーチを加味して選択したパラメータ間の比較を

Table 3. Overall properties about WBCD data（AVE± S.D.）

	subjects		1	2	3	4	5	6	7	8	9
train	benign	133	2.96±1.65	1.26±0.65	1.38±0.77	1.3±0.83	1.92±0.55	1.37±1.19	2.08±1.01	1.23±0.88	1.1±0.53
	malignant	67	6.78±2.66	6.28±2.79	6.45±2.53	5.37±3.03	5.37±2.36	8.1±2.88	6.31±2.33	5.76±3.37	2.1±2.08
test	benign	135	2.89±1.59	1.27±0.68	1.39±0.83	1.44±1.14	2.23±1.22	1.29±1.06	2.09±1.15	1.23±0.79	1.05±0.39
	malignant	65	7.23±2.36	6.86±2.65	7.03±2.52	5.49±3.12	5.28±2.51	7.54±3.26	6.29±2.29	5.89±3.37	2.69±2.85

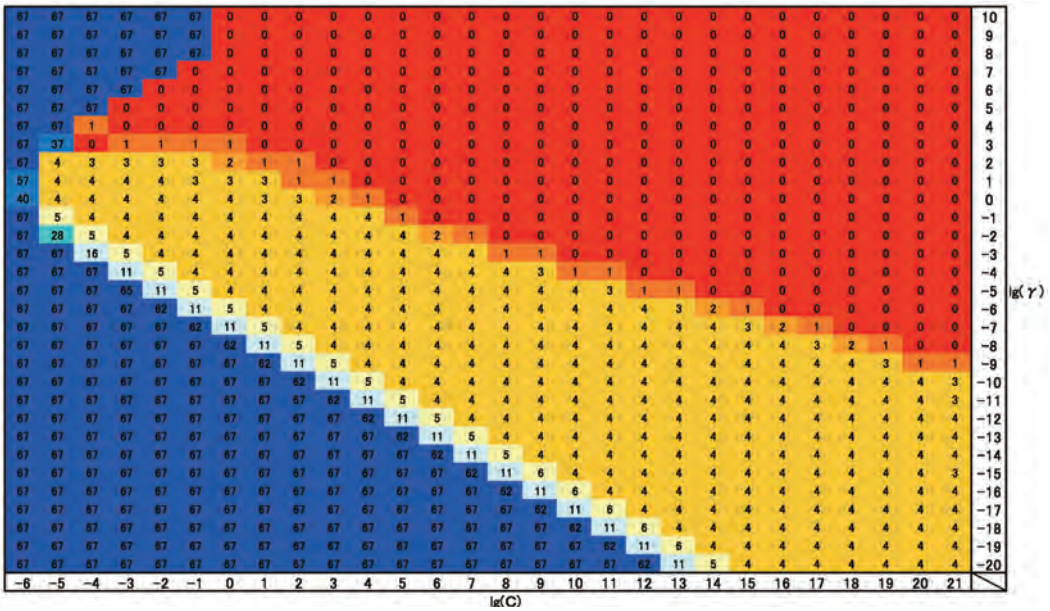


Fig. 3. The number of false negatives is denoted on each cell based on the grid-search

Table 4. The calculated correct rates (%) of the malignant (/benign) are given by BRNN and SVM for 200 test samples

model		correct rate（%） for malignant	correct rate（%） for benign
BRNN	400 MCMC steps	98.5	95.6
（mid12, ARD－）	800 MCMC steps	98.5	95.6
SVM （C－SVC, RBF）	C＝1, γ ＝1 Accuracy: 97% #FN: 4	96.9	96.3
	C＝4, γ ＝2 Accuracy: 97% #FN: 1	100	94.8

Table 5. The calculated correct rates（%）of hyperthyroid（/normal）obtained based on the LOO calculation with BRNN and SVM using 85 normal females and 49 hyperthyroid females are shown, for sets of 3 parameters（Alkaline Phosphatase, Serum Creatinine, Total Cholesterol）

model	correct rate（%） for hyperthyroidisms	correct rate（%） for normal
BRNN （mid12, ARD－, 400 MCMC steps）	90.0	96.5
SVM （C－SVC, RBF kernel, C＝4, γ ＝8）	91.8	97.6

見てみると，全体の正答率は変わらないものの，偽陰性が最も少ないパラメータ（C＝4, γ ＝2）ではテストデータの予測において悪性の正答率が100%となり，確実に判定することができた．この事からも，SVMでは適切にパラメータを設定することにより，BRNNと比べてより目的に沿った予測・判定を行わせることができるのではないかと考えられる．

血液の基本的検査データを用いた甲状腺機能異常者（バセドウ病患者）の予測

東北大学病院，およびJR仙台病院の甲状腺機能亢進症（バセドウ病）患者49名と健常者（対照群）85名の基本的検査値（ALP, S-Cr, TC）を属性データとして用いて，機能亢進症あるいは健常の2クラスを教師信号として与えてモデルを構築し，BRNNとSVMでの計算結果の相違をLeave-One-Out（LOO）法により確認した．LOO法とは，学習データの中から1つのデータを抜き取り，残りのデータを用いて学習し，その後抜き取ったデータについて予測テストを行うという手順を全サンプルに対して繰り返すやり方である．なお，計算にあたってSVMのパラメータについてはグリッドサーチを行って決定し，BRNNについてはこれまでの研究から妥当と思われる条件を採用した．Table 5に示された計算結果では，SVMではBRNNの場合と比べて健常者，亢進症患者ともに誤判定が1名ずつ減り，全体的に分類精度が向上している．

Fig. 4にBRNNとSVMのそれぞれについて，出力値（亢進症予測値）を横軸にとり，縦軸上部に健常群，縦軸下部に機能異常群と分けて度数分布を描いた図を示す．全体的な傾向として，BRNNでは推定率0%と100%に予測が凝集しやすい傾向が，すなわち，両極に分類しがちな性質がある．個別のデータを見てみると，BRNNでは亢進症推定率30%として偽陰性となっていた患者が，SVMでは出力値66%にて陽性と判定された．この患者の甲状腺ホルモン値（FT4，東北大学病院の基準値は，女性0.88～1.5 ng/dL）は3.58 ng/dLであり，臨床的には亢進症患者としてスクリーニングできれば有用と考えられるサンプルである．一方でBRNNでは41%と疑われていた患者がSVMでは22%にまで低下したケースがあるが，この患者のFT4値は2.20 ng/dLと低く，学習データの中でも3番目に低い値をもつサンプルであり，SVMでの判定結果はBRNNでの

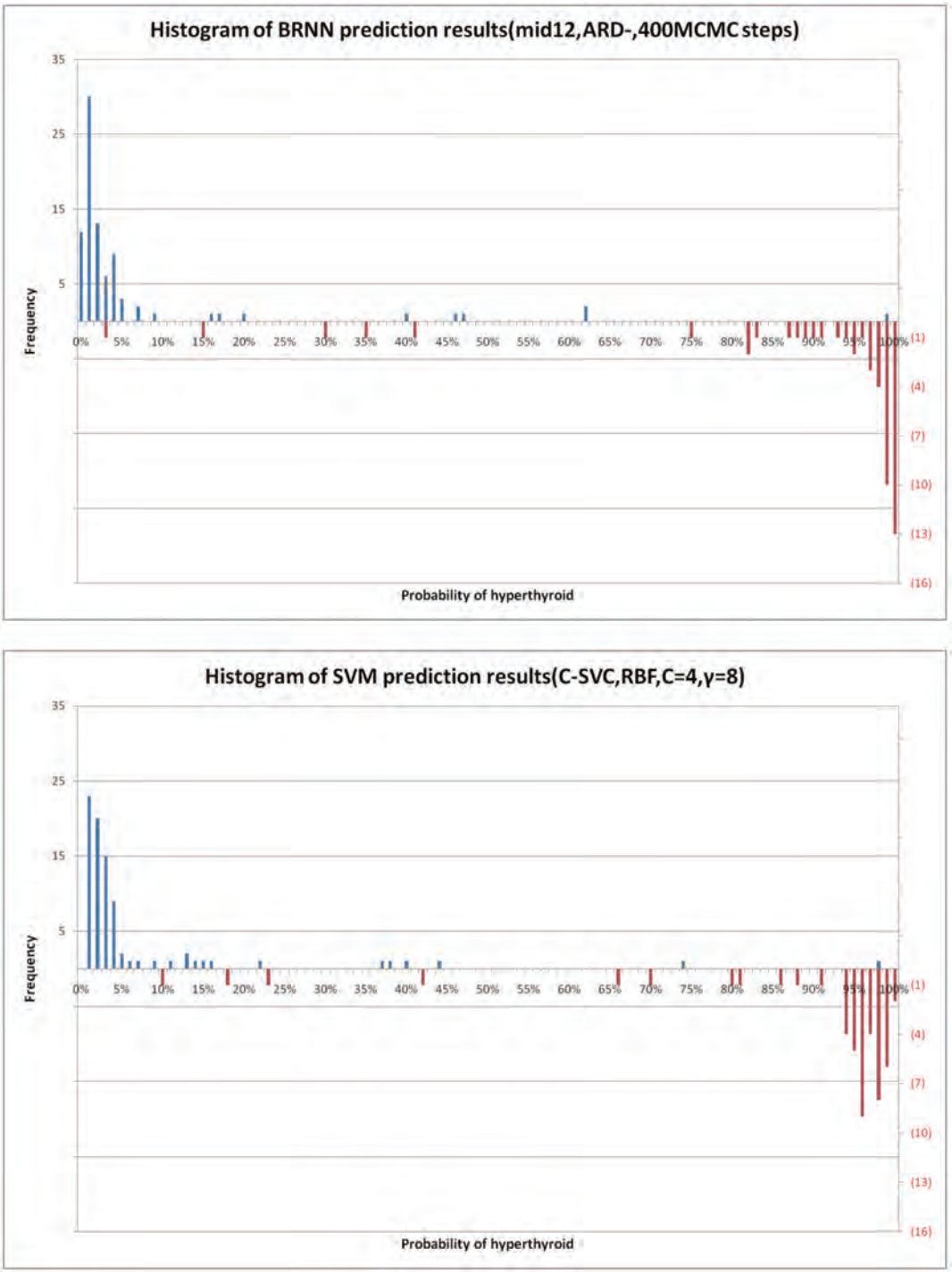


Fig. 4.　Details of the prediction are shown in histogram

Probability of hyperthyroid（%）is plotted on x-axis, while the number of samples（blue: normal females, red: hyperthyroid female patients）is plotted on y-axis.

判定結果より FT4 値との対応においても望ましいことが確認できた.

本論文では医療データの解析にサポートベクトルマシン（SVM）が有用な手法となりうるかを検討した. SVM の“マージン最大化”に基づく優れた分類能力を生かすためには RBF カーネルを採用して，グリッドサーチ法による最適パラメータの決定が大切である.

Wisconsin Breast Cancer Data の分類や甲状腺機能異常者のスクリーニングといった医療データの解析においては学習データの分類において偽陰性が最も生じにくいという条件もパラメータ決定に重要であり，そうすれば，SVM は判別分析よりは大きく分類精度が向上，BRNN 型ニューラルネット

ワークよりは特に2群の境界近くに位置するデータの分類において精度が向上することが明らかとなった．従って，健診施設や病院において甲状腺機能異常者のスクリーニングを行う場合に，我々がこれまで使用してきた2種類のニューラルネットワークに加えて，SVMも有用な解析手法として活用できるものと期待できる．

REFERENCES

1）MacKay D. J. C., *Neural Computation*, **4**, 448－472 (1992).

2）Neal R. M., "Bayesian Learning for Neural Networks," Springer, New York, 1996.

3）Kohonen T., "Self-Organizing Maps," Springer, Berlin, 2000.

4）Hoshi K., Kawakami J., Sato W., Sato K., Sugawara A., Saito Y., Yoshida K., *Chem. Pharm. Bull.*, **54**, 1162－1169 (2006).

5）Sato W., Hoshi K., Kawakami J., Sato K., Sugawara A., Saito Y., Yoshida K., *Biomed. Pharmacother.*, **64**, 7－15 (2010).

6）Aoki S., Nakatsuka K., Hoshi K., Kawakami J., Sato K., Sato W., Sugawara A., Saito Y., Sato K., Yoshida K., *The 129th Annual Meeting of the Pharmaceutical Society of Japan* (2009).

7）Sato K., Aoki S., Nakatsuka K., Hoshi K., Kawakami J., Sato K., Sato W., Sugawara A., Saito Y., Yoshida K., *The 50th Scientific Meeting of Japan Society of Ningen Dock* (2009).

8）Vapnik V., "The Nature of Statistical Learning Theory," Springer, New York, 1995.

9）Hsu C-W., Chang C-C., Lin C-J., Technical report, Department of Computer Science and Information Engineering, National Taiwan University (2003).

10）Chang C-C., Lin C-J., "LIBSVM: a Library for Support Vector Machines," http://www.csie.ntu.edu.tw/~cjlin/libsvm/

11）Neal R. M., "Software for Flexible Bayesian Modeling and Markov Chain Sampling," http://www.cs.toronto.edu/~radford/fbm.software.html

12）Keerthi S. S., Lin C-J., *Neural Computation*, **15**, 1667－1689 (2003).

13）Lin H-T., Lin C-J., Technical report, Department of Computer Science and Information Engineering, National Taiwan University (2003).

14）Mangasarian O. L., Wolberg W. H., *SIAM News*, **23**, 1 & 18 (1990).

15）Wolberg W. H., Mangasarian O. L., *Proc. Natl. Acad. Sci. U.S.A.*, **87**, 9193－9196 (1990).