サポートベクターマシンに基づく医療データからの事例発見

鈴木　英之進，菅谷　信介，津本　周作

1. はじめに

近年，電子化されたデータの爆発的な増加と諸IT技術の急速な進展に伴い，大量データの知的解析が比較的低コストで可能となった．この状況を背景に，データマイニング，あるいはデータベースからの知識発見（Knowledge Discovery in Databases: KDD）と呼ばれる新しい研究分野が誕生した．KDDの目的は妥当性，新規性，および理解可能性を満たし，有用となる可能性があるパターンをデータから抽出することである[4]．KDDは，1990年代初頭に誕生して以来，学界と産業界の両方で多大な関心を集めている．

KDDは，計算機による学習機能の実現を目的とする機械学習を母体の一つとするが，より幅広い種類の問題を指向する[6]．1990年代半ばより，機械学習における多くの研究は，表形式データからのクラス予測モデルの導出である分類学習に矮小化してしまった．このような研究は，正答率を評価指標とすることからも分かるように予測を目的とする．これに対しKDDは有用性を評価指標とし，予測よりも幅広い（意志）決定を目的とする[6]．現在研究分野としてのKDDは黎明期にあり，形式化された問題枠組の中で有名手法の改良を提案することだけではなく，種々の知識発見問題を形式化して新しい手法を提案することも重要視されている．

事例発見は，表形式データからの有用な例集合の特定と定義され，KDDの中でも新しい発見問題と位置付けられる．KDDにおけるパターン抽出は，主に分類学習，ルール発見，およびクラスタリングが用いられてきた．これらにおいて発見されるパターンは，それぞれ上述のクラス予測モデル，属性間の制約を表すルール集合，および全例集合の分割であるクラスター集合である．事例発見はクラスタリングに類似するが，全例集合を分割するのではなく，有用なクラスターを少数個発見する点が異なる．事例発見は，意志決定においては興味深い例集合が限定されている場合が多いと考えられる上，2.1節で述べる種々の利点を持つため，今後注目を集めると予想される．

事例発見は，学習目的が異なる既存の学習手法で得られた情報に基づいて行うことも可能である．サポートベクターマシン（Support Vector Machine: SVM）は，クラス予測モデルとして最近傍例への距離を最大とする超平面を求める分類学習手法であり，属性数は極めて多いがノイズがない問題で高い正答率をあげている[2, 3]．ここで，SVMのクラス予測モデルが高い正答率をあげられるなら，知識発見の観点からも有用であると予想される．著者の菅谷と鈴木はこの素朴な発想を手法として実現し，実際の医療データへの適用と領域専門家である津本による評価を経てその有用性を確認した[8]．本論文は事例発見手法の提案と医療データを用いた検証であると同時に，SVMの分類学習問題とは異なる知識発見問題への応用事例ともなっている．

2. 提案手法

2.1 事例発見

例えば共通の疾患を持つ患者たちが，(属性-値) ペアを用いた表形式データとして表されているとする．ただし疾患の種類を表す属性がクラスとして指定され，クラスはウィルス性か細菌性のどちらかの値をとりうる．このデータから，ウィルス性や細菌性の疾患を持つ典型的な患者の集合や，もう片方の疾患と区別がつきにくい患者の集合を発見できれば，これら少数の患者を注意深く検討することにより疾患についての理解が深まると考えられる．事例集合は，前章で述べたクラス予測モデルやルールに比較すると，データを抽象

すずき　えいのしん，すがや　しんすけ
横浜国立大学工学部
〒240-8501 横浜市保土ヶ谷区常盤台 79-5
つもと　しゅうさく
島根医科大学医学部
〒693-8501 島根県出雲市塩冶町 89-1

化していないために一般性は低いが，具体的であるために理解しやすいと考えられる．この例では，具体的な患者の集合は疾患を決定する手続きや（属性-値）間の関係に比較して分かりやすく，領域の理解につながる仮説を生成するきっかけとなる場合があると予想される．ただし知識発見が対象とするデータでは通常属性が多いため，事例発見は重要属性を特定できることが望ましい．

事例発見を次のように形式化する．入力は次に示すnの事例から構成されるデータ集合である．

$$(\mathbf{x}_i, y_i) \quad \mathbf{x}_i \in \Re^m,\ i=1, 2, \cdots, n \tag{1}$$

ただし$\mathbf{x}_i$はクラスy_iに属するm次元ベクトルである．$\mathbf{x}_i$を記述するm個の属性を$a_1, a_2, \cdots, a_m$とする．出力は，K個の属性$b_1, b_2, \cdots, b_K$で表されたq個の事例集合$\Pi_1, \Pi_2, \cdots, \Pi_q$である．ただし出力される属性集合は元の属性集合を限定したものである（$\{b_1, b_2, \cdots, b_K\} \subset \{a_1, a_2, \cdots, a_m\}$）．出力される事例集合$\Pi_i$は元の事例集合を限定したものであり（$\Pi_i \subset \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$），各出力事例集合は互いに排反である（$i \neq j$なら$\Pi_i \cap \Pi_j = \phi$）．

事例発見の計算量は$O(2^m 2^n)$である．もっとも事例発見の真の難しさは，データに記述されてなくユーザが暗黙に思っている事例に関する特徴を特定することである．

2.2 サポートベクターマシン

ここでは提案手法で用いたサポートベクターマシン（SVM）を説明する．

2.2.1 適用問題

SVMは式(1)のデータにおいて2クラス分類問題を解決する．

$$y_i \in \{-1, 1\} \tag{2}$$

SVMはデータ集合から次式で示すm次元空間に存在する超平面を求める．

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \tag{3}$$

ここで，$\mathbf{w}$はm次元ベクトル，bは係数，$\mathbf{u} \cdot \mathbf{v}$は$\mathbf{u}$と$\mathbf{v}$の内積である．ここではCortesらの定式化[3]にしたがい，次式が成立するとする．

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i,\ i=1, 2, \cdots, n \tag{4}$$

ここで，$\xi_i (\geq 0)$は誤分類の程度を表すペナルティ係数である．この定式化は，2クラスが線形分離不可能である場合にも適用可能であるが，誤分類が比較的少ない場合に適する．誤分類が多い場合には通常カーネル[2, 3]が使用されるが，本論文では考えない．

図1にSVMの概念を示す．超平面L_+と超平面

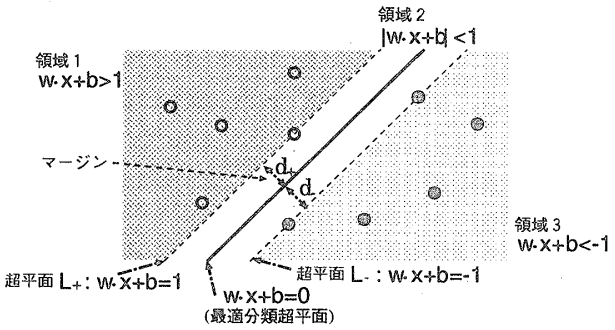

図1　サポートベクターマシンの概念図

L_-間の距離$\rho = d_+ + d_-$をマージンと呼ぶ．超平面(3)とベクトル$\mathbf{x}_0$の距離は$|\mathbf{w} \cdot \mathbf{x}_0 + b| / \|\mathbf{w}\|$で与えられることから，$\xi_i = 0$となる訓練例に対して，マージン$\rho$は$2/\|\mathbf{w}\|$となる．SVMが求めるクラス予測モデルを最適分類超平面と呼ぶ．2クラスが線形分離可能である場合には，最適分類超平面はρを最大にする超平面である．このとき，領域の境界線上にあるベクトルが最適分類超平面を決定する．この最適分類超平面を決定するベクトルをサポートベクトル（Support Vector：SV）と呼ぶ．また，領域1と領域3に分布し，それぞれのクラスに正しく分類されたベクトルを非サポートベクトル（NSV：Non-Support Vector）と呼ぶ．

2.2.2 最適分類超平面

誤分類事例を考慮して最適分類超平面を求めるためには，式(4)を条件とし，次の$\Phi(\mathbf{w}, \xi)$を最小化する最適化問題を解く．

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{5}$$

ただし，Cはユーザにより与えられるパラメータである．式(4)，(5)の最適化問題は2次計画問題の主問題であるためラグランジュ乗数α_iを用いて式(6)で与えられる双対問題を考えることで解くことができる[1]．

$$\begin{aligned} &\text{最大化} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{n}\alpha_i \\ &\text{条件} \quad 0 \leq \alpha_i \leq C,\ i=1, 2, \cdots, n \\ &\qquad \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned} \tag{6}$$

式(6)の最適解を$\bar{\alpha}_i$とすると，最適分類超平面$\bar{\mathbf{w}}\mathbf{x} + \bar{b} = 0$は次式で与えられる．

$$\begin{aligned} \bar{\mathbf{w}} &= \sum_{i=1}^{n}\bar{\alpha}_i \mathbf{x}_i y_i \\ \bar{b} &= -\frac{1}{2}\bar{\mathbf{w}} \cdot (\mathbf{x}_r + \mathbf{x}_s) \end{aligned} \tag{7}$$

ただしベクトル$\mathbf{x}_r$とベクトル$\mathbf{x}_s$はそれぞれ$y_r = 1$

と $y_s=-1$ のクラスに属する任意のSVである．

SVMにより，事例 $\mathbf{x}_i$ は $0<\bar{\alpha}_i<C$ を満たすSV，$\bar{\alpha}_i=0$ を満たすNSV，および $\bar{\alpha}_i=C$ を満たす誤分類事例に分類される．

2.2.3 解法

本論文ではSVMの解法として，Plattにより提案された順次最小最適化アルゴリズム（Sequential Minimal Optimization: SMO）[7]を用いる．SMOは式(6)の部分問題を順次解くことで解を求める高速アルゴリズムである．部分問題は2つのラグランジュ乗数から構成され，解くことが可能な最適化問題が選択される．SMOは各部分問題において選択された2つのラグランジュ乗数の解を求めていき，2次計画問題の目的関数を最小化していくことで解を得る．

2.3 サポートベクターマシンに基づく事例発見

2.2.2項で述べたようにSVMは最適分類超平面をクラス予測モデルとして出力し，事例をSV，NSV，および誤分類事例に分類する．われわれは出力事例集合 Π_1，Π_2，Π_3 を次のように定めた．

(1) Π_1：SV（他方のクラスとの境界事例）．

(2) Π_2：最適分類超平面から最も遠い H%のNSV（クラスの典型事例）．

(3) Π_3：誤分類事例（他方のクラスと紛らわしい事例）．

最適分類超平面の法線ベクトル $\bar{\mathbf{w}}=(w_1, w_2, \cdots, w_m)$ を用いた簡単な重要属性選択ヒューリスティックスを述べる．属性 i に関して各クラスの事例が極端に異なる分布をとらなければ，属性 i のクラス弁別力は $\bar{\mathbf{w}}$ における各成分の絶対値 $|w_i|$ で近似的に表される[8]．本手法はこの値が最も大きい K 個の属性を出力属性として選択する．ただし属性の単位を考慮しないと小さな値をとる属性が有利となってしまうので，属性 i についての平均値 x_m と分散 s を用いて，属性値 x_i を式 $z_i=(x_i-x_m)/s$ で z_i に正規化する．

3. 髄膜炎データベースへの適用

3.1 髄膜炎データベース

前章で提案した手法の有効性を評価するために，表1に示す髄膜炎データベース[9]を用いた．このデータは首都圏の某都市の指定3次救急指定病院において，昭和54年から平成4年までに入院治療した140人の髄膜脳炎（meningoencephalitis）患者に関するものである．髄膜炎データベースは欠落値が一つの属性だけに現れることからも分かるように，前処理をほぼすべて終えたデータに分類される．髄膜炎データベースは，いくつかのコンテストで使用された，データマイニングにおける標準データである．

表1 髄膜炎データ

(属性群の説明) 属性	説明
(個人情報)	
AGE:	年齢
SEX:	性別
(診断)	
DIAG:	元々の診断
Diag2:	DIAGをVIRUSとBACTERIAにグループ化したもの
(来院時の状態)	
COLD:	何日前から風邪様症状があったか．
HEADACHE:	何日前から頭痛があったか．
FEVER:	何日前から熱があったか．
NAUSEA:	何日前から吐き気があったか．
LOC:	何日前から意識障害があったか．
SEIZURE:	何日前から痙攣・てんかんがあったか．
ONSET:	発症様式
(来院時の身体検査)	
BT:	来院時体温
STIFF:	来院時項部硬直
KERNIG:	来院時 Kernig 徴候
LASEGUE:	来院時 Lasegue 徴候
GCS:	来院時 Grasgow Coma Scale (意識障害の状態を示すスコア)
LOC_DAT:	意識障害の状態
FOCAL:	巣症状
(来院時の精密検査)	
WBC:	白血球数
CRP:	炎症性蛋白
ESR:	血沈
CT_FIND:	CT所見
EEG_WAVE:	脳波所見
EEG_FOCUS:	脳波の局所異常の有無
(治療と予後)	
CSF_CELL:	髄液細胞数
Cell_Poly:	髄液中多核球数
Cell_Mono:	髄液中単核球数
CSF_PRO:	髄液蛋白
CSF_GLU:	髄液ブドウ糖
CULT_FIND:	培養・免疫検査で，原因菌・ウィルスが発見できたか．
CULTURE:	発見できた菌・ウィルス名
THERAPY2:	実際の治療
CSF_CELL3:	3日後の髄液細胞数
CSF_CELL7:	7日後の髄液細胞数
C_COURSE:	最終的な転帰
COURSE(Grouped):	C_COURSEを2つにグループ化したもの
RISK:	危険因子
RISK(Grouped):	RISKを2つにグループ化したもの

3.2 適用条件

2つの値をもつ属性の中から，Diag 2（診断属性を2つにグループ化したもの），EEG_WAVE（脳波所見），CT_FIND（CT所見），CULT_FIND（原因菌・ウィルスが発見の有無），COURSE（Grouped）（最終的な転帰を2つにグループ化したもの）およびRISK（Grouped）（危険因子を2つにグループ化したもの）の6属性をそれぞれクラスとして，6個の問題を設定した．これらは医学の専門家にとって興味がある問題となっている．

実験ではSVMが連続値属性を対象とすることから，3個以上の値を持つ名目属性を削除した．以降COURSE（Grouped）およびRISK（Grouped）を，それぞれCOURSEおよびRISKと省略する．

ここでCOURSEのような治療内容や治療後の経過に関する属性は，診断過程後に得られる値である．診断時の属性であるDiag 2，EEG_WAVE，CT_FIND，およびCULT_FINDについての問題に関しては，これらの属性をその時点では入手できないとして除外した．同様にRISKを全問題から削除した．

さらに各属性において，属性値の分散が5を超える値は異常値と考えその値を除く最大値より大きな値に変更した．パラメータは$K=H=10$，$C=10^5$と設定した．Cをこの値に設定したのは誤分類事例が少ないモデルを生成するためだが，問題によっては計算時間が長くなってしまう．計算時間が24時間を超えても分類モデルが求まらない場合は，Cの値を10^4，10^3とこの順に変更することにした．このためCの値は，RISKでは10^4，CT_FIND，CULT_FIND，およびCOURSEでは10^3となった．

なお提案手法のSVMをフィッシャーの線形判別(Fisher's Linear Discriminant: FLD)[5]に代えたものを比較手法とした．

4. 実験結果

本章では，発見結果を医学的知識に照らし合わせて提案手法の有効性を検証する．

4.1 重要属性の選択に関する結果

最初にDiag 2をクラスとした問題で正規化の有効性を検証した．正規化を用いない場合には比較的大きな値をとるCell_PolyとCSF_CELLは選択されず，用いる場合には選択された．これらの属性は，医学的に見て重要であることが分かっている．このことより，この場合正規化は有効だと考えられる．

表2 提案手法/素朴手法/FLDに基づく手法についての，属性選択に関する平均スコア

Task	属性選択		
	妥当性	意外性	有用性
Diag2	**4.8**/4.6/4.7	**1.0**/1.4/**1.0**	**4.8**/4.5/**4.8**
EEG_WAVE	**4.7**/4.0/4.2	**1.2**/2.2/1.3	4.1/4.1/**4.5**
CT_FIND	**4.9**/4.8/4.8	**1.0**/**1.0**/1.5	**5.0**/4.5/**5.0**
CULT_FIND	**4.8**/4.7/4.7	1.8/**1.4**/1.6	**4.9**/4.8/4.8
COURSE	4.9/**5.0**/4.5	1.2/**1.0**/1.5	4.9/**5.0**/4.7
RISK	**4.8**/4.1/4.5	**1.1**/1.9/1.5	**4.9**/4.4/4.6

妥当性と有用性について最も大きなスコアを太字で示し，意外性については最も小さなスコアを太字で示す

次に提案する重要属性選択方法の有効性を評価した．津本は発見結果を次の基準にしたがい1（最低）から5（最高）までの5段階で評価した．

(1) 妥当性：発見結果が医学的な知識と一致する度合．

(2) 意外性：発見結果が医学的な知識と部分的には一致するが，説明がつかない度合．

(3) 有用性：発見結果が医学的に有用な度合．

本実験の比較手法は，前章で述べたFLDに基づく手法と，次に述べる素朴手法を用いる．後者は，属性の重要度を例空間においてその属性に垂直な超平面で正しく分類できる事例数で評価する方法である．表2に，これら3つの手法を6個の問題に適用した結果を示す．

表より本手法は他の2つの手法に比べて総合的には優れていることが分かる．妥当性は6問中5個の問題において他の手法より良い結果であることが分かる．また有用性は6問中4個の問題において他の手法より良い．意外性に関しては，本手法は3つの手法の中で比較的低いスコアになっている．これは本手法が属性選択に関して保守的であることを示している．

実験において素朴手法は，属性間の依存関係を考慮していないために評価が低いと考えられる．本手法は属性数が多く訓練例集合にノイズがない問題に適する．一方FLDに基づく手法は，クラス予測モデルが境界例ではなく訓練例の分散に基づく方が適切な問題に適する．次節で明らかとなるが，FLDは2クラスの線形分離が難しいCOURSEに関しては本手法よりも総合的には優れている．

4.2 事例発見に関する結果

表3に典型事例発見の結果を示す．各列はそれぞれDiag 2，EEG_WAVE，CT_FIND，CULT_FIND，COURSE，およびRISKをクラスとした問題を表す．

表3　典型事例の発見に関する結果

Task	Di	E	CT	CU	CO	R
expert	31	35	28	20	25	29
SVID	13/14	11/15	10/12	10/10	9/12	9/13
FLD	11/15	11/16	10/12	8/12	11/13	9/12

表4　境界事例の発見に関する結果

Task	Di	E	CT	CU	CO	R
expert	34	35	49	58	47	39
SVID	23/24	21/29	22/26	28/31	19/32	26/32
FLD	9/15	6/13	7/12	7/12	11/12	10/14

表5　誤分類事例の発見に関する結果

Task	Di	E	CT	CU	CO	R
expert	0	9	15	36	27	22
SVID	-/0	-/0	15/20	36/58	19/32	8/10
FLD	0/7	8/12	10/18	29/35	11/31	10/16

また“expert”は専門家によって選択された事例数を表し，“SVID”および“FLD”は提案手法とFLDに基づく手法における（専門家と一致する事例数)/（発見事例数）を表す．この表より，各手法が専門家の事例を部分的にしか発見できないことが分かる．これは，各手法が2クラスの分類問題に関する情報だけに基づくのに対し，専門家は他の情報も用いるためだと考えられる．本手法はCOURSEを除きFLDに基づく手法とほぼ同等であるか優れている．本手法がCOURSEで劣るのは，後述するように2クラスの線形分離が困難であることが関係すると考えられる．

表4に境界事例発見の結果を示す．提案手法はCOURSE以外の問題においてはFLDに基づく手法に比較して優れている．これは，この問題においてはSVMで得られるSVが知識発見においても意味があることを示している．

表5に誤分類事例の発見結果を示す．この表より，いくつかの問題は専門家にとっても難しいことが分かる．表のSVIDにおける誤分類事例数より，各問題における2クラスについての線形分離の難しさが分かる．CULT_FINDとCOURSEは，それぞれ58，32事例が誤分類され，2クラスの線形分離が難しいことが分かる．このことは，COURSEにおいて提案手法がFLDに基づく手法に劣っている主な理由であると推測される．その他の問題については，提案手法はFLDに基づく手法に比較して専門家の判断と似ていることが分かる．この実験において2クラスの線形分離が容易な問題については，本手法は有効だと考えられる．

表6　COURSEの事例集合における，LOC_DATとFOCALに関する事例数の分布

事例の種類	共に-	片方は+
-の典型	23	2
-の境界	30	17
+	6	35
-	70	29

以上より本手法はFLDに基づく手法に比較して，2クラスの線形分離が比較的容易である場合には，境界事例発見と誤分類事例発見において領域専門家の判断と類似する．これは前節で述べたように，訓練例のノイズが少なく，高次元の分類問題にSVMが適しているためだと考えられる．

5.　事例発見の効用

津本は提案手法で発見された事例を基に，いくつかの興味深い知識を発見した．例えば彼はCOURSEが一の典型事例と境界事例を比較するうちに，LOC_DATとFOCALの組合せが他よりも重要であるとの知見を発見した．発見の根拠となった事例数の分布を表6に示す．この発見は，LOC_DATが脳全体の機能を表しFOCALが脳の局所的な機能を表すことから説明できる．この知見は医学的知識から予想できるが，強い傾向ではないために統計的手法では発見できなかった．この発見は，提案手法が典型事例と境界事例を発見し，主要属性だけについて事例を表示したために可能となった．

6.　おわりに

本稿ではサポートベクターマシン（SVM）の事例発見への応用を提案した．この手法は各クラスの典型事例，境界事例，および誤分類事例を発見し，これらを重要でない属性を省いて出力する．われわれは知識発見の共通問題として使用されている医療データを用いて，提案手法の有効性を医学的観点から評価した．提案手法は線形判別手法と素朴手法に比較して，重要属性の選択に優れていた．また線形判別手法に比較して，2クラスの線形分離が比較的容易である場合には，境界事例と誤分類事例の発見に関しては領域専門家の判断に類似することが分かった．

知識発見は主にクラス予測モデルやルール集合などの高次パターンを求める問題を対象としてきたが，これからは事例集合という低次パターンを求める事例発

見も重要になると考えられる．例えばルールは通常，前提部に少数の属性しか含まないので情報が少なく，結論部のクラスだけしか考慮しない．これに対して事例は，より多くの属性を含むので情報が多く，典型事例と非典型事例を比較することで重要な発見につながることがある．ただし事例の属性数が極端に多い場合には可読性が低い．提案手法は重要属性を選択するためこの問題を解決しており，津本による種々の発見に貢献したと考えられる．

参考文献

[1] M. S. Bazaraa and C. M. Shetty: *Nonlinear Programming*, Wiley, New York, 1979.

[2] C. J. C. Burges: "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121-167, 1998.

[3] C. Cortes and V. Vapnik: "Support Vector Network", *Machine Learning*, Vol. 20, No. 3, pp. 1-25, 1995.

[4] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth: "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*, pp. 1-34, AAAI/MIT Press, Menlo Park, Calif., 1996.

[5] D. J. Hand: *Discrimination and Classification*, Wiley, New York, 1981.

[6] T. M. Mitchell: "Machine Learning and Data Mining", *CACM*, Vol. 42, No. 11, pp. 31-36, 1999 (邦訳：機械学習とデータマイニング，*CACM* 日本語版，Vol. 1, No. 1, pp. 7-12, 2000.).

[7] J. Platt: "Fast Training of Support Vector Machines Using Sequential Minimal Optimization", B. Schölkopf, C. Burges, and A. J. Smola (eds.), *Advances in Kernel Methods—Support Vector Learning*, pp. 185-208, MIT Press, Cambridge, Mass., 1998.

[8] S. Sugaya, E. Suzuki, and S. Tsumoto: "Instance Selection Based on Support Vector Machine for Knowledge Discovery in Medical Database", *Instance Selection and Construction for A Data Mining*, pp. 395-412, Kluwer, Norwell, Mass.

[9] 津本周作：「知識発見手法の比較と評価のための共通データ」，人工知能学会誌，Vol. 15, No. 5, pp. 751-758, 2000.(データは http://www.slab.dnj.ynu.ac.jp/challenge2000/より入手可能.)