文字認識と svm 法

Historical view of svm method for character recognition

安田道夫　中島由美　北村浩治　二階堂真理恵

Mitio Yasuda Yumi Nakashima Koji Kitamura Marie nikaido

要旨

svm 法が OCR（文字認識装置）で実用に供されるようになったのは１９９０年代半ばからであるが，その手書き文字認識 性能の高さ (正読率と処理速度) から，一部では文字認識（個別文字の）の技術は完成し，今後はさまざまな学習機械への応用があるのみと言う受け止め方をされているようである．しかし，文字サンプルの中からその母集合の認識に有効なサンプルを選択する目的には，類似度を用いるほうが計算も簡単で，選択したサンプルの有効性も svm 法によるそれと変らない．svm 法を特徴づけているカーネルや対判定も，類似度法では不用である．また，パーセプトロン形の多層ニューロンモデルや多層ニューロンモデルやベイズ確率は有用だろうか？ 本稿では両者の類似性とと相異点について，一部模式的実験をまじえて考察する．

1　はじめに

svm 法は，古来１９７０年代半ばから試みられて来たが，そのアプローチの一つである線形連立一次不等式を解く必要性が制約となっていたが，１９９０年代前半に，カーネルトリックを導入することで，線形連立一次不等式を非線形化し，対判定と云う制約はあるもの，ニューロンモデルにおける各種パラメータを計算し，解を求めることが可能になった．

2　手書文字認識における svm 法の効果 [1],[2]

svm 法は任意の未知カテゴリのパターン **x** をニューロンの入力とし，その出力を適当な閾値を通すことで，各ニューロンごとに想定する，二個のカテゴリの何れであるかの判断を **y** として出力する，いわゆる対判定 **pairwize decision** 方式を用いる．

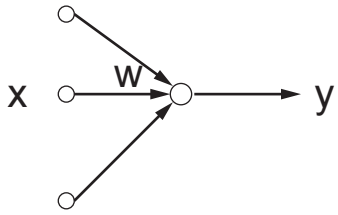

図 1　svm 法のニューロンモデル

svm 法では，図 1 に示すニューロンを想定し，その入力ベクトル **x** と出力 **y** の関係を式 1 が成立するように求める．

$$\begin{aligned} y &= \mathrm{sign}(\mathbf{w}^T x - h) \\ &= \mathrm{sign}\left(\sum_{i \in s} \alpha_i \cdot t_i (x_i^T x - h)\right) \end{aligned} \tag{1}$$

式 1 で，しきい値 $h > 0$，および入力ベクトル **x**，重みベクトルの各成分と α_i はすべて非負 (≥ 0) とする．式 1 を，仮説 $H_+ : \alpha_i \cdot t_i (x_i^T x_+ - h) = \Delta_+ \ (\Delta_+ > 0)$ と $H_- : \alpha_i \cdot t_i (x_i T x_- - h) = -\Delta_- \ (\Delta_- > 0)$ で記憶すべき $\alpha_i (> 0)$ と，x_i を選択して，仮説 H_+ と H_- に対応する２枚の超平面を求められる．

ニューロンモデルにもとづく svm マシンを構成するための基本的な手順を説明したが，svm マシンを実際に適用するには，ソフトマージンとカーネルの導入が必須とされる．

2.1　ソフトマージン

文字認識への適用を考えると，まず認識すべき文字のサンプルを学習サンプルと対照サンプルの二群に分割し，学習サンプルの中からサポートベクトルとして使用するサンプルを残す．この作業は式 1 を満たすことを判断の基準とするが，入力ベクトル x に対する出力 y が，$\Delta_+ > y > -\Delta_-$ になる場合，すなわち幅 $\Delta_+, -\Delta_-$ の区間になる場合がある．このような場合は α_i を十分に大きくしても避けられるが，それではマージンを確保して，選択するサポートベクトルの数を減らす目的にはそぐわない．そこで h の値を学習サンプル x_i ごとに微調整するパラメータ ξ_i を導入して，

$$h \Rightarrow h - \xi_i \quad (\xi_i \geq 0) \tag{2}$$

とする．実際には，さらにつぎの式 3 に示すパラメータ γ と ν_i を導入してよりこまかな調整を可能にしている．

$$\alpha_i = \gamma - \nu_i \tag{3}$$

ソフトマージン法における三つのパラメータ ξ, γ, ν を導入することで，対判定の前提ではあるが，線型分離ではない場合でも有効な識別のための超平面を定めることができる．ソフトマージン法の直感的な意義は，すべての学習ベクトル（の頂点）を二つの平行な超平面のいずれかにおくのではなく，それらの中間や外側にも置くことで，超平面上の学習ベクトル（サポートベクトル）の数を減らす効果もある．

2.2　カーネルトリック

識別問題が複雑で，識別面が本質的に平面よりなんらかの曲面の方が適している場合，学習サンプル（ベクトル）を非線型関数で別の空間に写像し，その空間で線形識別を行なう**カーネルトリック**と呼ばれる方法がある．

カーネルトリックは，特徴ベクトル x を非線形の写像 $\phi(x)$ で変換し，この新たな空間で線形識別を行なう方法で，本質的に非線形識別が適している問題など，より複雑な問題への適応能力が高まるとされる．

svm 法では各種のパラメータを求める計算は，すべて二つの特徴ベクトル x_1 と x_2 の内積だけに依存する．

$$x_1^T x_2 \quad \Rightarrow \quad \phi(x_1)^T \phi(x_2) = K(x_1, x_2) \tag{4}$$

式 4 の $K(x_1, x_2)$ をカーネルと云う．カーネルは計算回数が多くなりがちなので，できるだけ簡単な形が望ましいとされる．

カーネルの例

多項式カーネル　$K(x_1, x_2) = (1 + x_1\ Tx_2)^p$ (5)

G*auss* **カーネル**　$K(x_1, x_2) = \exp\big(-\frac{(\|x_1 - x_2\|^2}{2\sigma^2}\big)$ (6)

シグモイドカーネル　$K(x_1, x_2) = (\tanh x_1^T x_2 - h)$ (7)

2.3　svm 法のふるまい

svm 法の特長は，対象とする学習サンプルすべてを正しく識別するように，システムのパラメータ・・・というよりデータ＝サンプル・・・を選択する（できる）ことである．ここで識別用データとして選択さ

表 2.3svm 法による **etl6** 数字部分の認識結果

					(%)		(秒)
サベ	特長成分	認識対象	認識対象文字数	選択 sv 数	誤読数	正読率	認識時間
ポク	スカラ	奇数	6910	960	23	99.67	4
｜ト		偶数	6910	1015	16	99.77	4
トル	１６成分	奇数	6910	813	8	99.84	30
法	ベクトル	偶数	6910	744	7	99.90	30
	スカラ	奇数	6910	1	55	99.25	178
摂		偶数	6910	1	56	99.23	178
動	１６成分	奇数	6910	1	23	99.67	180
法	ベクトル	偶数	6910	1	22	99.68	180

れる学習用サンプルの学習サンプル全体に対する割り合いは，1990 年頃とされる当時の常識からするときわめて大きいが，その後のマイクロプロセッサを中心とするハードウェア処理性能の急速な高速化（少くとも 10^3) と相まって，正読率と処理速度が向上した．svm 式の構成原理や解釈については，現在でもさまざまな問題点や誤解があると思うが（これらについては別章で触れる），結果としてこの方式で達成された知見の歴史的意義は大きい．

図 2.3 に手書文字データベース **etl6** の手書数字部分を対象にした，svm 法による認識結果と併わせて摂動法による処理結果を示す．

光学的文字認識（**OCR**) ではとくに正読率が重視されるので，摂動法でもスカラ，および１６成分特徴の双方で９９％台の正読率ではあるものの，svm の方が半桁程度すぐれている．また，処理速度の差はさらに大きく１桁以上速い，

3　類似度と距離

類似度と距離は相互に密接な関係があり，適切な処理のもとで，一方から他方を導くことができる．以下にまずそのために用いる簡単な数学的手段について述べる

3.1　ヘッセの標準形式

ヘッセの標準形式とは，任意の次元の多次元空間における超平面の式を

$$w^T x = \text{定数} \tag{8}$$

とするとき，つぎの式をヘッセの標準形と云う．ここで右辺は定数であれば良い[*1]

$$d = \frac{w^T(x - x_*)}{\sqrt{w^T w}} \tag{9}$$

ここで，x_* は超平面上の任意の点の座標またはベクトルを，d は多次元空間の任意の点 x からこの超平面に降ろした垂線の長さ（距離）を符合を含めてあらわす．

つぎに，$w, x,$ および x_* を，$w^T w = 1, x^T x = 1, x_*^T x_* = 1$ に正規化しておくと，式 9 は，つぎの式 10 の形になる．

$$d = w^T(x - x_*) \tag{10}$$

文字認識への適用を考える場合，列ベクトル x，および x_* の各成分は

$$x_i \geq 0,\ x_{*,i} \geq 0 \qquad (i = 1, 2, \cdots, I) \tag{11}$$

としてよい．さらに，$w \equiv x_*$ とすると，式 10 は式 12 の形になる．

$$d_{i,*} = x_*^T(x_i - x_*) = x_*^T x_i - x_*^T x_* = s_{*,i} - 1 \leq 0 \tag{12}$$

図 2 にヘッセの標準形による距離と類似度の関係を示す．なお図では式 12 の添字 x_i を $x_\circ$ と描いた．

式 12 の $d_{i,*}$ は $d_{i,*} \leq 0$ だが，

$$s_{i,*} + |d_{i,*}| = 1 \tag{13}$$

式 13 の関係が成立する．

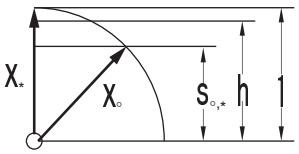

図 2　ヘッセの標準形における距離と類似度の関係

3.2　類似度モデル

サポートベクトルを選択するための式 1 右辺の括弧内の部分を

$$x_i^T x - h \Rightarrow x_i^{mT} x_j^n - h = s_{i,j}^{m,n} - h \tag{14}$$

[*1] １９５５年ころ（昭和２０年代）には高校の教科書に記載されていた

と書くことにする．ここで x_i^m は第ｍカテゴリの第ｉサンプル，x_j^n は第ｎカテゴリの第ｊサンプルを意味する．

$$s_{i,i}^{n,n} - h = 1 - h \tag{15}$$

$$s_{\substack{i,j \\ j\neq i}}^{m,m} - h \geq \max_{\substack{m\neq n \\ 1\leq i\leq I}} s_{i,j}^{m,n} \tag{16}$$

$$s_{\substack{i,j \\ j\neq i}}^{n,n} - h + \xi_i^n \geq \max_{\substack{m\neq n \\ 1\leq i\leq I}} s_{i,j}^{m,n} \tag{17}$$

ただし　$\xi_i^n = \frac{1}{2}\left(1 - \max_{\substack{m\neq n \\ 1\leq i\leq I}} s_{i,j}^{m,n}\right)$ (18)

学習サンプルの中から，学習サンプルの全てを正しく認識するために必要な参照サンプルを選択する，この作業は一意的なものではなく，学習サンプル全てを選択しても良いし，適当な規則の元で全ての学習サンプルの類似度を参照して参照サンプルを選択しても良い．

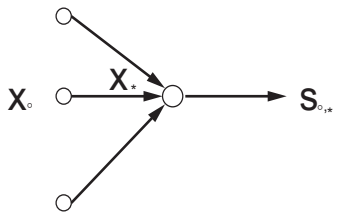

図３　類似度モデル

図３に類似度を利用するニューロン形の演算記憶要素の例を示す．x_* が選択された任意の学習サンプルを表わす，類似度は早い時期（１９７０年代）からＯＣＲに利用されて来た．また，二次元相関法と云う意味ではそのＯＣＲへの利用はさらに１０年さかのぼる１９６０年代から利用されていた [3]．

類似度行列を利用する認識参照サンプルの選択例

認識参照サンプルを選ぶ方法はいくつか考えられるが，何れの場合も，原則として学習サンプルを全て正しく認識するように，認識参照サンプルを選択する．つぎに，そのような選択方法の例を述べる．

1. 第 m カテゴリの第 i サンプル（x_i^m）を参照サンプルに撰んだとき式 15 が成立し，正しく認識される．したがって，全ての学習サンプルを認識参照サンプルとして選択することにすれば，当然のことながら全ての学習サンプルを正しく認識する（図４（ａ））．
2. 学習サンプル x_i^m について式 16 が成立するとき，すなわちある学習サンプルと，同一カテゴリの他の学習サンプルの類似度が他カテゴリの全ての学習サンプルとの類似度より大きいとき，この学習サンプルを認識参照サンプルから除外する侯補として一時的にしるしをつけておき，実際に除外するか如何かは同一カテゴリの他の全ての学習サンプルを調べてから判断する

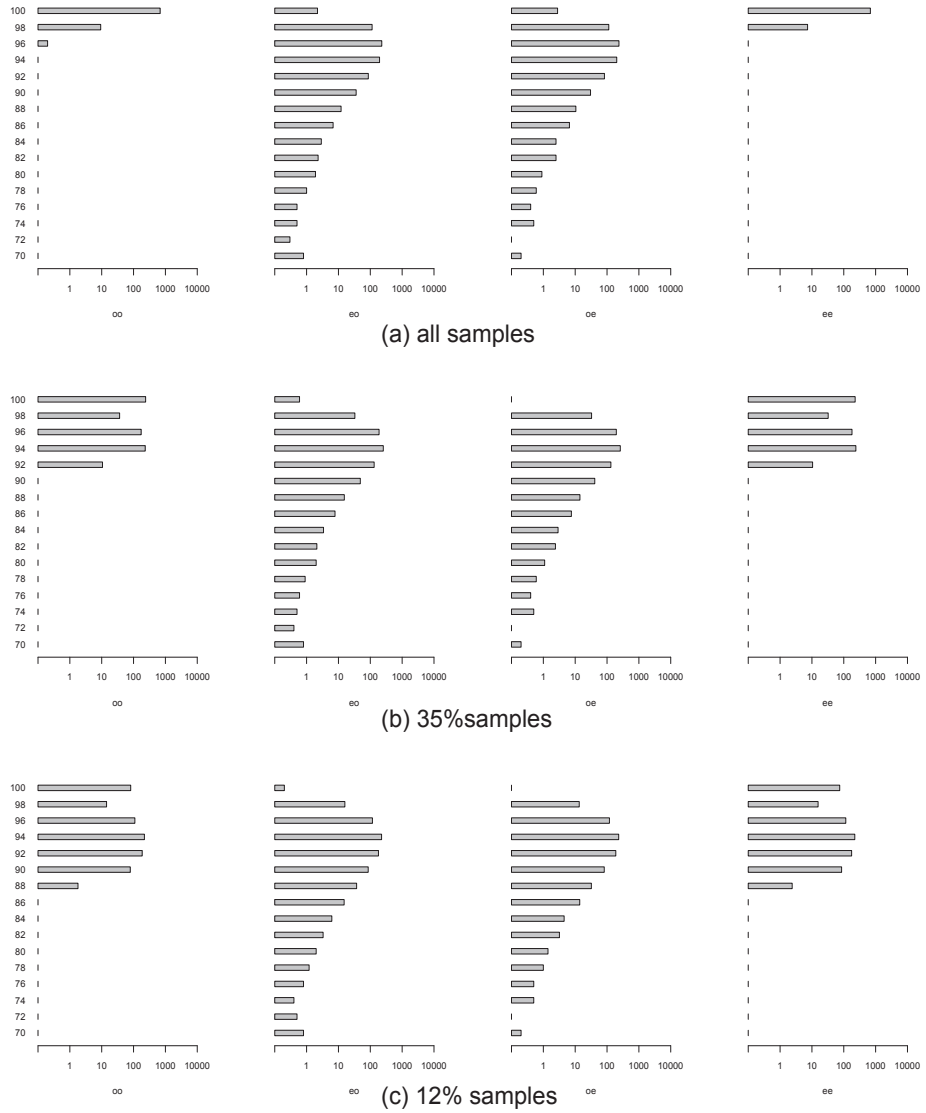

図 4　認識参照サンプル選択比率と類似度

3. 類似度の閾値に相当する h の値を低めに設定（0.9⇒0.8〜0.85）する代わりに，すべての他カテゴリの学習サンプルとのあいだの最大類似度とのマージンを指定する（式 17）

以上の２と３の処理で正しく学習サンプルは，それ自身を認識参照サンプルとして選択する（式 15）.

3.3　距離モデル

式 1 の括弧内の部分をヘッセの標準形式を用いて書き直す．

$$\begin{aligned} x_i^T x_* - h &= x_i^T x_* - 1 + 1 - h \\ &= x_i^T (x_* - x_i) + 1 - h \\ &= (s_{i,*} - 1) + (1 - h) = -(1 - s_{i,*}) + (1 - h) = d_{i,*} + (1 - h) \quad (19) \\ &= s_{i,*} - h \quad (20) \end{aligned}$$

式 13 などで示したように $0 \geq d_{i,*} \geq -1$ だから，式 1 の α_i に相当する定数を $\dfrac{1}{(1-h)}$

$$1 \geq \text{式 19 の右辺} \geq -\frac{h}{1-h} \quad (21)$$

また，$d_{i,*}$ の変動範囲を $0 \geq d_{i,*} \geq -(1-h)$ に限れば

$$1 \geq \text{式 19 の右辺} \geq 0 \quad (22)$$

式 22 の形になる．これは式 1 の仮説 $H_1 : t_i = 1$ に対応するが，これを明示するため，$x_* \Rightarrow x_+$ として書き直すと

$$H_1 : 1 \geq \alpha_i \cdot t_i (x_i^T x_+ - h) \geq 0 \Rightarrow 1 \geq \alpha_i (x_i^T x_+ - h) \geq 0 \quad (23)$$

$$H_2 : 1 \leq \alpha_i \cdot t_i (x_i^T x_- - h) \leq 0 \Rightarrow 1 \geq \alpha_i (x_i^T x_- - h) \geq 0 \quad (24)$$

式 23 と式 24 は，類似度モデルの式 15 と等価である．ソフトウェアマージンの設定には多少の調整は必要だろうが．

式 19〜式 24 を用いて示したように，通常の類似度とこの場合の距離は形式・量とも同等と見なすことが出来る（図 5）．なお，距離で求めたサポートベクトル [2] を用いていわゆる単純類似度法でその効果を確認した．

その逆の確認実験は行なっていない．

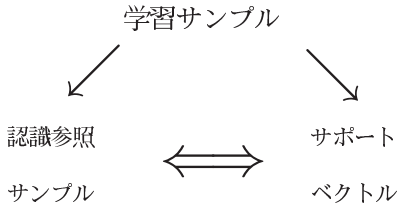

図 5　学習結果の互換性

また一般に類似度を用いる場合，対判定方式をあらわに用いる事はないが，これは類似度が順序集合であり対判定の要請を自ずと満たしているからである．

4　svm 法の役割

パターン認識，とくにＯＣＲ（光学的文字認識）は，１９９０年ころ，ローゼンブラット流のニューロンモデル（パーセプトロン）のパラメータを決定するための線形連立不等式を解く問題と，本質的に線形独立の制約を有するこの問題を解く手段としての非線形写像した空間で解くことにより，線形の呪縛を脱し…１パターン当りのニューロンの入力端子数＝連立方程式の変数の数に拘束されずに…多数のサンプルを利用できるようになった結果，現在では個別文字認識についてこれ以上やるべき事はないと云う向きもあるようだ．

歴史的事象として svm 法が果たした役割は確かに大きかったであろうが，半面無用の処理や，誤った幻想を与えていないだろうか？ましてや，文字認識を人の知的活動の一つの核心と捉えるなら，現在の個別文字認識技術はまだ科学的と言うより技巧的処理に依存している面が大きいのではないだろうか？以下，思いつくままに順不同で列挙する．

svm 法は何故１９９０年ころに花開いたか？＝＝＝＞コンピュータの高速化が最大の理由

このころは国産各社のスーパーコンピュータが出そろい，つぎのマルチコアの数〜十数 GFLOPS 時代を控えていた．半面，パソコンは無論のこと，ミニコンもせいぜい十〜数十 MIPS の性能で（現在のパソコンの方が 10^4〜10^5 速い），特別のプロジェクトで，時間と人を注ぎ込むような時以外はシミュレーションの手段としては利用されなかった．一方半導体の集積化と高速化はそこそこ進んでおり，ミニコンシミュレーションの数十〜百倍程度の処理速度を実現することは容易だった．

対判定は必要か?===＞ 使わない方がよい

筆者の一人は**対判定**と云う用語の提案者（の一人?）だが，**対判定**と云う熟語自体は統計的検定論の分野で普通に使用される術語である．文字認識で云う**対判定**は，対判定で判定する二つの仮説が共通部分を持たないことである．これは論理的には切れ味が鋭いが，実用面で便利かと云うとそうではない．また類似度（距離も測定方法が定まっていればそうだが）のように大小だけで判定できる量の方が汎用的である．対判定は筆蹟鑑定や，あーでもないこうでもないと鳩首協議の場合にむいている．

サポートベクトルは何をサポートしているのだろうか ===>針か旗棹では？

ヘッセの標準形式形の意味では類似度＝１のベクトルあるいは図形関数（グラフ) に正規化した後では一個しか存在しない，サポートと云う術語はシュワルツの**超関数の理論**に出て来るが，その意味は全く別である．また手書文字はカテゴリ間に明確な境界は期待できない（図 6 参照)．

比較的文字の質が良いとされる図４の手書文字データベース etl6 の対照サンプルの認識結果は９９．９％を維持しているもの，類似度７０～９０％のサンプルは十分に多い．このようないわば隙間に他カテゴリの文字が入り込む怖れは常にあるだろう．

カーネルは必要か

カーネルはパーセプトロン形式のニューロンモデルに，学習サンプルの一部（あるいは全て）を，学習サンプルを全て正しく識別すると言う意味で有効かつ選択的に記憶させる事がことができると云う意味で必要な道具ではあるだろう．しかし，学習サンプルからその全てを正しく識別するために行こなうサンプルを選択する目的は，学習サンプル間の類似度を比較すれば達成できる（図 5 参照)．カーネルのもう一つの役割とされる特徴空間の写像もその実質的な効果が超関数の理論で定義された**畳み込み（convolution）**[4] と同等であれば，類似度の計算に組みこむことも出来る．

いつまでパーセプトロンか

1963 年頃，当時の通産省電気試験所がパーセプトロンの評価装置を外部企業に試作させたと云うことを，20 年程後に聴いたことがある．その当時からパーセプトロンには，ある種の線形連立方程式、あるいは不等式を解く機能があることは判っていたが，そのころの関心は逐次学習，レイヤー間のランダム結合による何か好いことないかな的願望も多く，その後も永く手掛けられて来た．しかし，どういうアプローチだったにせよ，それをきわめて効果的な svm 法に到達した先駆者（筆者は知らない）には敬意を表すが，それは同時に多くの誤解も生じた．実際，1970 年ころには，米国の Recognition Equipment 社が”RETINA” と言う名称で高速のアナログ式相関器を多数そなえたＯＣＲを出荷しており，その計算能力で例えば **etl6** の手書き数字部分に適用すれば，毎秒数千文字の認識速度を達成できただろう．

ここ数年，最速のスーパーコンピュータの特長は，並列化した多数の **GPU** をアーキテクチャに取り入れたことだと言う．ＯＣＲの場合図 3 に相当する **GPU** は単純な積和乗算器で良いから，ニューロンモデルとコンピュータアーキテクチャの対応も良い．

5　今後の課題

現在ＯＣＲは技術的に完成の域に達し，半面強固な壁にぶつかっていると云う説が支配的に見える．しかし，ＯＣＲ開発当初からの課題のなかでほぼ解決されたように見えるのは，個別の印刷・手書文字を高速かつある程度の精度での処理が実現した，あるいは実現のめどがたったというだけで，個別文字認識本質的に内在する限界や，文字の質感など，工業製品としての課題にもほとんど手がついていない．ましてや，文字認識と人の知能との係わり，あるいは文字認識の機能と脳との生物物理的関連付などにはほとんど手がついていない．

カラテオドリの定理

図６と７にカラテオドリ（Charatheodory)[5],[6] の定理の概念図と文字認識に促した模式図を示す．

カラテオドリの定理の結論は単純明解で，等角写像と云ういわば性質が良い写像で図に示したように変形できる事を示している．

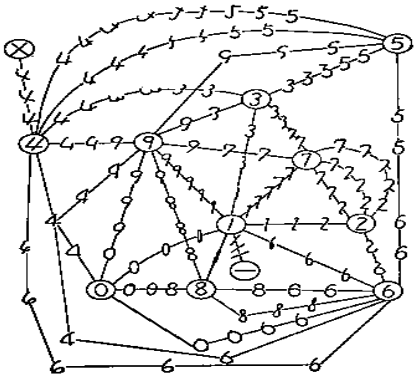

図６　手書数字の変形模式図

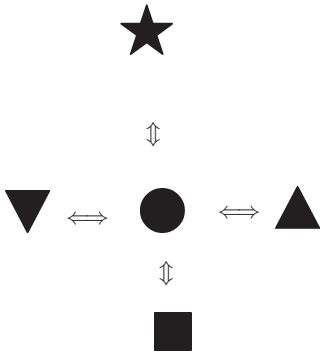

図７　カラテオドリの定理の概念図

人の知能とは何だろう

svm の成果を見ると生物は記憶さえすれば利巧になるのかと思うかもしれない．しかし，京都大学霊長類研究所によると，十年かけて教育したオランウータンに尤もらしく云えばある種の観察力と記憶力のテストをしたところ，オランウータンの方が，実験に協力した京大生より２倍以上すぐれていたとのことだ．

似たような話は，スコットランドヤードが雇用している，事件現場スケッチの異能者が BBC で紹介さ

れたことがある．それではシャーロックホームズは？

文字認識の生体的モデル

これは空想的，あるいは幻想的な呟きに類するものだが，パターン認識や文字認識には，何かと云えば多次元ベクトル，超平面等が登場する．もっとも単純と看做す人もいる類似度法（実際に，わざわざ単純類似度法と呼ぶ人もいる）も，超関数そのものである．三次元モデルでさえ様々なグラフに描かれているが，ときとして，本質を見え難くしている場合もある。

多変数関数は必要かと云うヒルベルトの第１３問題に対するコルモゴロフの解は [7]，複数の簡単な 1 変数関数を加えると云う簡単な 2 変数関数があれば良いというものだそうだし，超関数の世界では関数とのアナロジーが成立する場合があるそうだから，文字認識を処理するシステムも，なによりも三次元世界の住人であるわれわれの目と頭の中で実現されているのだから，判り易い単純なモデル化が可能ではないだろうか？

そこで，地球の中心から表面にかけてある立体角で切り出したようなものを仮定し類似度 1 ＝距離 0 は地球表面とする・・・・・

参考文献

[1] 栗田多喜夫；”http://home.hiroshima-u.ac.jp/tkurita/lecture/svm/index.html”

[2] Chih-Chung Chang and Chih-Jen Lin, LIBSVM :”a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm

[3] G.L.Fischer,etal edited,”OPTICAL CHARACTER RECOGNITION”,1962,Spartan Books

[4] 岩村聯邦訳，” 超関数の理論”，岩波書店，1959

[5] 安田、中野、藤本，” 手書き文字の認識技術”，信学誌，1978，2

[6] 辻正次，” 複素変数関数論”，共立出版，

[7] 国沢，梅垣，” 情報理論の進歩”，岩波書店，1965