文字認識における類似度法と svm 法
Similarity method and svm method for character recognition

北村浩治　二階堂真理恵　中島由美　安田道夫
Koji Kitamura Marie nikaido Yumi Nakashima Michio Yasuda

要旨

文字認識，とくに個別文字認識に適用する場合，両者は学習サンプルから一定の手順で，認識処理に使用する参照サンプルを選択する学習過程を経て，具体的なパターン，すなわち個別文字の認識能力を獲得する．どちらの方法も同じ学習サンプルから参照パターンを選択する教師付き学習過程とみなせる．各々の方法で選択するパターンは，必ずしも一致しないし，各々の方法の中でも学習の手順に応じて異りうる．また，**svm** 法はその歴史的経緯からニューロンモデル上で学習サンプルが二個一対の異なるカテゴリのどちらに属するかを判断するための超平面をもとめる方法への拘りがあるため，対判定をもちいている． この場合カテゴリ間の境界をきめ細かに設定できる反面，学習パターンは母集団そのものではないので，このようなきめ細かさが逆効果になることも考えられる．また **svm** 法では学習パターンが超平面のどちら側にあるかを判定するための内積計算の回数が，類似度の場合は参照パターンの総数 N の程度であるのに，対判定の場合には $N^2/2$ 回必要になる．類似度法は 1970 年代初頭から印刷文字の認識に広く利用されており，手書き文字認識に利用するには参照パターンの選択方法を検討すれば良いことになる．

1　はじめに

文字認識装置（**OCR**）は 1950 年代初頭に実用化され，当初の特殊字体を対象とするものから，通常字体，複数字体 (**multifont**)，全字体 (**omnifont**) を処理するものへと，1960 年代末までに進化した．これらの **OCR** は何れも印刷文字を対象としたものであり，**全字体**と言う語句は象徴的な表現である．しかし，この時期の末頃発表出荷された **Recognition Equipment** 社の **retina** は，最大 256 種の異る字体の英文タイプライタ（一字体当り，英大小文字・数字・記号を含む約 100 字種）で印字された文書の処理が可能と謳っていた．**retina** は，最大 25,600 個の参照パターンを記憶できるので，各カテゴリの参照パターン数を固定とせず可変にできるとすると，例えば手書き数字の認識を高精度で行なうことも，原理的には可能だったと言うことになる，

類似度法と **svm** 法は，すでに論じたように [1]，認識に使用する参照パターンを学習サンプルの中から選択する点では同一だが，類似度法ではパターンを二次元関数とみなして，特定カテゴリの任意の学習サンプル g が少くとも一つの参照サンプル f に対して，適当に定める閾値 Δ より小さくないように参照パターンを選択する．ここで類似度 $\mathcal{S}_{f,g}$ は，式 (1) をみたすものとする．

$$0 \leq \mathcal{S}_{f,g} \leq 1 \tag{1}$$

類似度法でのこのような参照パターンの選択は，学習サンプル中の他カテゴリのサンプルを考慮せず実行できるので，参照パターン選択の手数はカテゴリ数と比例する程度になる．

一方，**svm** 法の場合はパターンを多次元ベクトルとし，異るカテゴリに属する学習サンプルを正しく分離する複数の超平面を決定する参照パターンをカテゴリ対ごと選択する，このため，参照サンプルを選択するための手数は，カテゴリ数の二乗の程度になる反面，異るカテゴリ間の境界をよりきめ細かに設定できるように見える．

しかし，個別文字，とくに手書き文字の場合，その変形量は大きく，異カテゴリ間の境界を定めることは本質的に困難である．また，かりに学習サンプルについては完璧な境界を定め得たとしても，対照サンプル（未学習サンプル）の認識への有効性は疑問である．

以上に述べたように，個別文字，とくに手書き文字の認識には本質的な限界があるが，当面認識に利用する参照パターンの個数を最大化すると同時に，その記憶に必要なデータ量と処理手数を最小化する必要がある．また，平均パターンを参照パターンに繰り入れたり，摂動法の採用もある程度の有効性を期待できる．

2　個別文字認識のための手書き文字データベース

文字認識システムの性能評価を行うにあたり，電子総合研究所 (現　産業技術総合研究所) の ETL6 と NIST Special Database 19 の 2 種類のデータベースから手書き数字データを使用する．本稿のデータについての説明は，すべて手書き数字の個別文字サンプルについてである．

ETL6 の個別文字は，64×63(横 $\times$ 縦) の領域に 4bit(16 階調) の濃度レベルであらわされる画像データである．データは，各文字につき 1,383 データが収録されている．使用にあたっては，簡単なノイズ除去処理により観測ノイズを取り除いてから，それを 2 値化したものを用いる．

NIST Special Database 19 は，米国国立標準技術研究所 (NIST: National Institute of Standards and Technology) によって提供されているデータベースである．NIST はさまざまなデータベースを提供しているが，この Special Database 19(SD19) は，手書き文字を収録したものである．個別文字は，辺々 128 の領域に 2 値であらわされた画像データを圧縮して格納している．ここに収録されているデータは，その収集時期などの違いによっていくつかのシリーズに分かれており，各シリーズの先頭は hsf という文字列を冠してある．データの収録

数は，どのシリーズでも各個別文字とも 5000 件以上を占めているが，少ないものでは，4500 件程度となっている．

表 1　NIST SD19　個別文字 (手書き数字) のデータ件数

category	hsf_0	hsf_1	hsf_2	hsf_3	hsf_4	hsf_6	hsf_7
0	5534	5472	5352	6613	5560	5939	5893
1	6008	5972	5816	6976	6655	6710	6567
2	5321	5263	5187	6360	5888	6086	5967
3	5592	5517	5505	6558	5819	6085	6036
4	5114	5188	5097	6150	5722	6010	5873
5	4603	4644	4566	5732	5539	5838	5684
6	5236	5284	5206	6402	5858	6051	5900
7	5589	5549	5459	6611	6097	6334	6254
8	5262	5244	5203	6320	5695	5966	5889
9	5190	5179	5076	6174	5813	6075	6026

3　単純平均による標準パターンの作成

素朴で直感的な手法である相関法では，認識対象の図形 (未知パターン) と各カテゴリを代表する図形 (標準パターン) との類似度 (相関係数) を求め，類似度が最大となるものをそのカテゴリと識別する．類似度 $\mathcal{S}$ は式 (2) によって求めることができる．

$$\mathcal{S}_{f,g} = \frac{(f,g)}{\|f\| \cdot \|g\|} \tag{2}$$

ここで，$\|f\|^2 = \sum_{x,y} f(x,y)^2$，$\|g\|^2 = \sum_{x,y} g(x,y)^2$，および $(f,g) = \sum_{x,y} f(x,y) \cdot g(x,y)$ である．

シミュレーション実験をするにあたって，データを学習用サンプルと未知サンプル (認識用データ) とに分ける．各カテゴリのデータを先頭から $1, 2, \cdots$ と付番して，偶数番目と奇数番目に分け，一方を学習用サンプル，もう一方を未知デサンプルとした．

各カテゴリの文字 (手書き数字の “0”–“9”) を代表する標準パターンを 1 カテゴリにつき 1 つの図形であらわすことにした場合の，ひとつの簡単な方法として，学習用サンプルを足し合わせ，平均値によって得られる標準パターンを作成したものをまず考える．

ETL6 では，各カテゴリに 1,383 個のサンプルがあるので，最後の 1 つを捨て，偶数と奇数の組を同数の 691 個ずつのサンプルにする．個々のサンプルは縦方向，横方向それぞれのヒストグラムをとって文字部分だけを矩形領域として切り出したデータにする．ただし，“1” のよ

図 1　ETL6 の単純平均による標準パターン例 (右：偶数，左：奇数)

図 2　NIST の単純平均による標準パターン例 (hsf_0) (右：偶数，左：奇数)

うに横方向の広がりが極端に狭い場合には，大きさの正規化によって形が大きく変わらないように例外処理を施す．切り出したデータは，20×20 の領域に大きさが 14×14 になるように整える．この整えたデータを足し合わせ，その平均値をとり濃度をそろえると単純平均の標準パターン (平均パターン) になる．NIST では，収録とカテゴリによりサンプル数がまちまちなので，使用するサンプルは，どのカテゴリも先頭から選んだ 4560 個に統一する．サンプルを偶数と奇数の組に分けた，2280 個のサンプルから 1 つのカテゴリの標準パターンを作成する．

NIST による標準パターンの作成処理は，使用するサンプルのデータ数をのぞき，ETL6 と同様である．ETL6 および NIST から作成した単純平均の例を図 1 と図 2 とで示す．

4　類似度法による参照パターンの選択方法

前節で述べた，単純平均による標準パターンは，各カテゴリを代表するデータが，カテゴリごとに 1 つだけになるようにしたが，実際の文字 (特に手書き文字) のひろがり (変形パターン) は，ほぼ無限と言ってもよく，また，あるカテゴリの文字と別のカテゴリの文字との境界も曖昧である．参照するパターンを 1 カテゴリにつき 1 つとせず，参照パターンとして実在のすべてを採用して記憶することはほぼ不可能であり，実用のためには，参照パターンを選択する採用基準を設ける必要がある．参照パターンを選ぶ基準はいくつが考えられるが，どの場合でも学習用データを正しく認識するように参照パターンを選択するものである．類似度法による参照パターンの選択方法は以下のように行う．

- 前提条件として採用する基準の類似度を閾値として設定．
- 各カテゴリの最初の参照パターンを適当に設定．(たとえば単純平均，学習データの 1 つめなど)
- 学習サンプルと同一カテゴリの既存の参照パターンとの類似度を求め，最大類似度が採用基準の閾値を満たさない場合，この学習サンプルを参照パターンに追加する．

類似度は，式 (1) のように 0 から 1 の間の値をとるので，閾値を 1 にした場合は，学習用サンプルに完全な同一データがない限り，すべての学習サンプルが参照パターンとして選択されるようになる．

表 2　類似度法による参照パターンの選択 (ETL6)

類似度の閾値	偶数の学習サンプル		奇数の学習サンプル	
	参照パターン数 (率)	自己正読率	参照パターン数 (率)	自己正読率
1.00	6910 (100.00)	100.00	6910 (100.00)	100.00
0.99	6446 (93.29)	100.00	6473 (93.68)	100.00
0.98	5693 (82.39)	100.00	5746 (83.36)	100.00
0.97	4332 (62.69)	100.00	4471 (64.70)	100.00
0.96	3141 (45.46)	100.00	3250 (47.03)	100.00
0.95	2268 (32.82)	100.00	2396 (34.67)	100.00
0.94	1691 (24.47)	99.99	1782 (25.79)	100.00
0.93	1291 (18.68)	99.99	1366 (19.77)	99.99
0.92	987 (14.28)	100.00	1077 (15.59)	99.97
0.91	801 (11.59)	100.00	871 (12.60)	99.93
0.90	652 (9.44)	99.97	708 (10.25)	99.93

表 3　類似度法による参照パターンの選択 (NIST hsf_0)

類似度の閾値	偶数の学習サンプル		奇数の学習サンプル	
	参照パターン数 (率)	自己正読率	参照パターン数 (率)	自己正読率
1.00	22800 (100.00)	100.00	22800 (100.00)	100.00
0.99	22150 (97.15)	100.00	22132 (97.07)	100.00
0.98	19999 (87.71)	100.00	19914 (87.34)	100.00
0.97	16707 (73.28)	99.99	16664 (73.09)	99.99
0.96	13330 (58.46)	99.96	13296 (58.32)	99.94
0.95	10331 (45.22)	99.78	10327 (45.29)	99.84
0.94	7973 (34.97)	99.57	7908 (34.68)	99.67
0.93	6136 (26.91)	99.21	6091 (26.71)	99.39
0.92	4673 (20.50)	98.67	4657 (20.43)	99.00
0.91	3638 (15.96)	97.93	3658 (16.04)	98.56
0.90	2898 (12.71)	97.63	2860 (12.54)	97.94

表 2 と表 3 に，学習サンプルから参照パターンを選択したときの，参照パターン数 (率)，その参照パターンと学習サンプルとの相関法による認識結果を自己認識率として示す．参照パターンとして採用する基準の類似度は 1.00 から 0.90 まで，0.1 刻みにした．なお，各カテゴリの最初の参照パターンとして単純平均の標準パターンを使用したが，表中の参照パターンの数に，この数は含んでいない．

5　svm 法によるサポートベクタ (sv) の決定方法

svm 法では，異なるカテゴリの学習サンプルが超平面によって分離できるとする．このカテゴリ間を完全に分離できる超平面は無数に存在するが，最適な識別面は，2 つのカテゴリ「真ん中」を通るものを求めるものであり，2 つのカテゴリの関係によって決定する．svm では，この超平面のまわりにある学習サンプルをサポートベクタ (sv) として選択する．svm 法では，これを学習モデルと呼んでいる．もともと 2 つカテゴリを分ける svm を多カテゴリの svm を適用する場合は，n カテゴリの問題を $n(n-1)/2$ 個のカテゴリ対からなる 2 カテゴリ問題に変換して，$n(n-1)/2$ の対判定となる．作成にあたっては，LIBSVM[3] を使用した．

表 4　svm によるサポートベクタの選択 (ETL6)

偶数の学習データ		奇数の学習データ	
参照パターン数 (率)	自己正読率	参照パターン数 (率)	自己正読率
960 (13.89)	100.00	1015 (14.69)	100.00

6　類似度法により選択された参照パターンによる認識シミュレーション実験

既に述べた「類似度法による参照パターンの選択方法」を使って作成した参照パターンと，未知データとで認識シミュレーション実験を行った．表 5 および表 6 にその実験結果を示す．ここで使用した参照パターンは，表 2 と表 3 の参照パターンであり，表 2 は表 5 と 表 3 は表 6 と，それぞれ閾値で紐付けられる．

表 5　類似度法による認識シミュレーション結果 (ETL6)

参照サンプルの閾値	学習：偶数 / 未知：奇数		学習：奇数 / 未知：偶数	
	誤読数	正読率 (%)	誤読数	正読率 (%)
1.00	26	99.62	25	99.64
0.99	26	99.62	25	99.64
0.98	26	99.62	25	99.64
0.97	26	99.26	24	99.65
0.96	26	99.62	22	99.68
0.95	27	99.61	24	99.64
0.94	25	99.64	23	99.67
0.93	30	99.57	27	99.61
0.92	27	99.61	24	99.65
0.91	32	99.54	26	99.62
0.90	36	99.48	28	99.59

表 6　類似度法による認識シミュレーション結果 (NIST)

参照サンプル	学習：偶数 / 未知：奇数		学習：奇数 / 未知：偶数	
の閾値	誤読数	正読率 (%)	誤読数	正読率 (%)
1.00	302	98.68	319	98.60
0.99	302	98.68	318	98.61
0.98	305	98.66	311	98.64
0.97	314	98.62	303	98.67
0.96	313	98.63	329	98.56
0.95	352	98.46	349	98.47
0.94	431	98.11	393	98.28
0.93	525	97.70	443	98.06
0.92	609	97.33	524	97.70
0.91	766	96.64	624	97.26
0.90	888	96.11	733	96.77

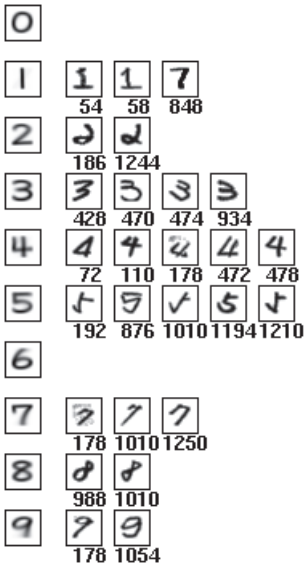

図 3 oe の誤読サンプル (ETL6)

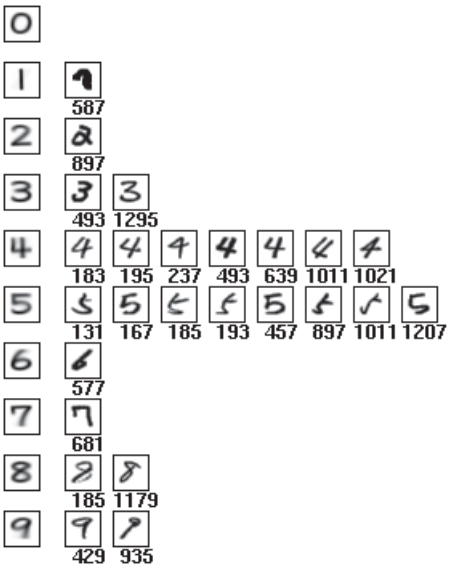

図 4 eo の誤読サンプル (ETL6)

ETL6 を使ってシミュレーション実験を行った結果，誤読となった文字サンプルの例を図 3，図 4 で示す．左端に番号なしで表示してあるのは，標準パターンである．番号が添えられてある文字図形は，誤読の文字サンプルで，添えられている番号は，データベースに格納されているデータに付番したものである．図 3 は学習サンプルに奇数 (o) データ，未知サンプルに

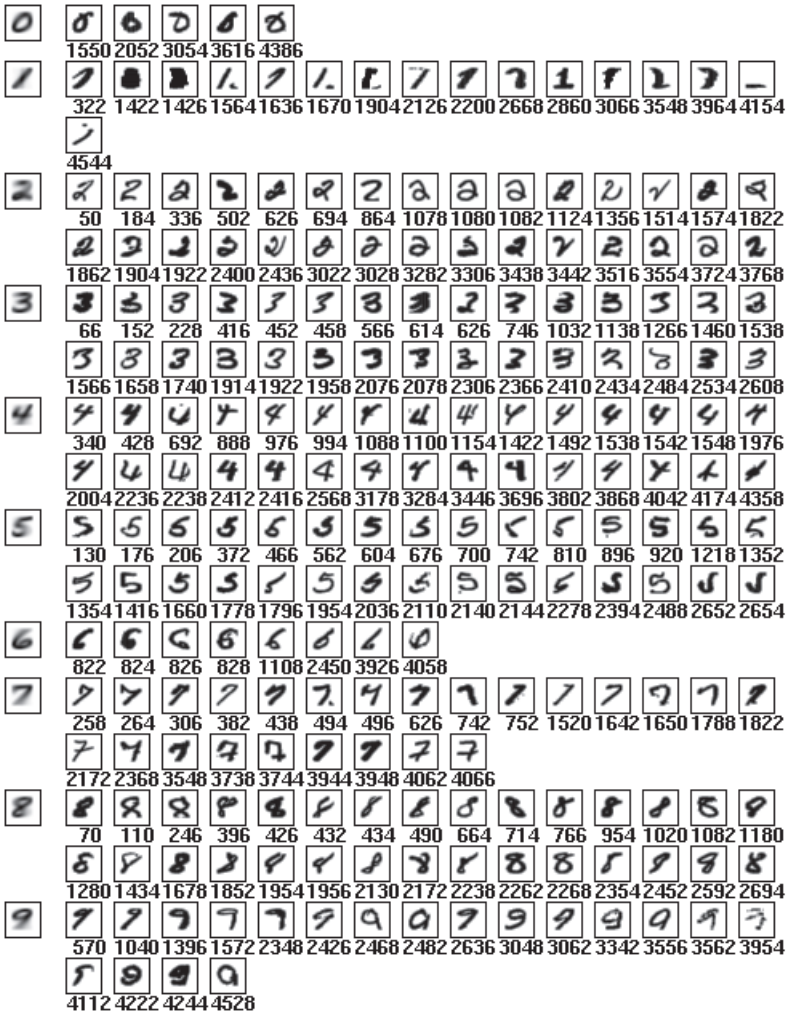

図 5　oe の誤読サンプル (NIST)(一部)

偶数 (e) データを使用したもので，これをここでは，“oe” と呼ぶことにする．図 4 は， eo での誤読サンプルを示したものである．どちらも，1 回の認識シミュレーションで誤読となった全サンプルを掲載している．

NIST を使ってシミュレーション実験を行った結果，誤読となった文字サンプルの例を図 5 で示す．誤読となったサンプル数が 1 カテゴリ 30 を越えるものは，紙面の都合により，サンプル数を 30 個までにしている．

7　類似度法と svm 法

ETL6 を使い，svm 法の学習によって選択された，(学習モデルの) サポートベクタ (sv) は，前出の表 4 の通りで，学習サンプル数の約 14% 程度である．svm 法のサポートベクタと，類似度法の参照サンプルが，同程度のものは閾値が 0.92 の辺りである．「svm 法」の場合と「閾値 0.92 の類似度法」の場合それぞれで得られたサポートベクタ (sv) 数または参照サンプル数の割合，認識率を図 6 で表す．便利のため図中および以下，サポートベクタも参照サンプルも sv と呼ぶことにする．円の左半分は sv の割合で，右半分は認識の割合である．

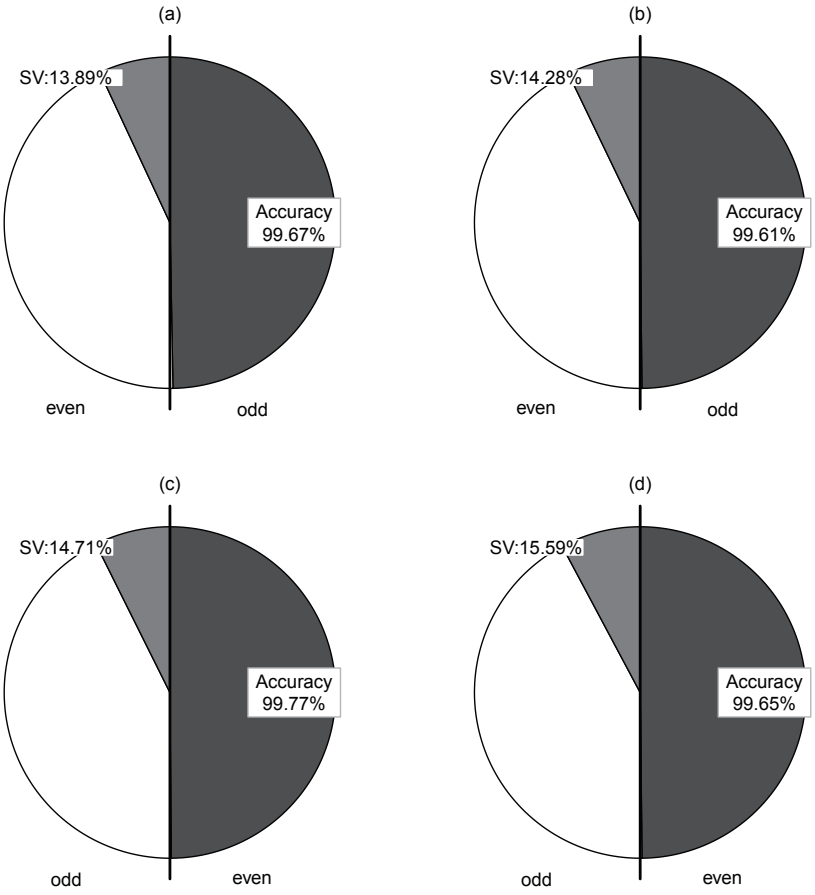

図 6　類似度法と svm 法の比較
(a)svm 法 (eo)，　(b) 類似度法 (eo)，　(c)svm 法 (oe)，　(d) 類似度法 (eo)

eo の場合，svm 法では sv が 960 サンプル，誤読が 23 サンプルとなり，類似度法では sv が 987 サンプル，誤読が 27 サンプルとなった．oe の場合，svm 法では sv が 1015 サンプル，誤読が 16 サンプルで，類似度法では sv が 1077 サンプル，誤読が 24 サンプルとなった．

8　考察

文字認識で一般的に有効と考えられる摂動法や方向性の特徴抽出などを用いなくとも標準パターンを 1 つとせずに，参照パターンを複数もつことにより ETL6 では 99% 以上の認識性能が得られることが分かった．NIST を使った認識率は，ETL6 を使った場合よりも認識率が低下したが，図 5 の誤読サンプルで見ても分かるように，判別が難しい文字サンプルが含まれているようである．

svm 法も，類似度による参照パターン選択法も，教師つき学習の機械学習であるが，類似度法による参照パターン選択法は，参照パターンの選択を同一カテゴリで決定しているのに対し，svm 法は，学習モデルの作成を他カテゴリとの関係から決定していることが決定的な違いである．svm 法についての解釈は様々ある [2, 3, 4] が，この方式は，他カテゴリとの分離面を決定するための sv を選んでおり，他カテゴリの増減によってその関係が変わってしまい，まったく異なる学習モデルができることになる．一方で，参照パターン選択法では，他カテゴリによって選択される参照パターンが変わることはない．ただ，どちらの方式をとったとしても，未知パターンは，学習パターンに含まれているわけではなく，各カテゴリ間には，連続的な変形パターンが無数に存在するので，未知のサンプルに対して必ずうまくいくものでもない．また，学習サンプルに本来は判別不能な文字サンプルや間違った文字サンプルが含まれた場合は，本来，認識不能な文字を認識できてしまうという不都合もある．これらについては，今後の課題である．

参考文献

[1] 安田，中島，北村，二階堂，” 文字認識と svm 法”，pp.33-43,vol.20, 明星大学情報学部紀要，2012

[2] 栗田，”http://home.hiroshima-u.ac.jp/tkurita/lecture/svm/index.html”

[3] 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

[4] 甘利，麻生，津田，村田，” パターン認識と学習の統計学”，岩波書店，2003