類似度行列を用いる学習サンプルの選択法
The method of selecting learning sample pattern with similarity matrix

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

要旨

類似度行列を用いて，学習サンプル集合から複数のサンプルを標準パターンとして適用する手法を提示する．本論文では，手書き文字データベース **ETL6** の手書き数字部分を用いた学習サンプルの選択および認識のシミュレーション実験を行ない，その結果について報告する．相関法に用いる特徴場としては，16 成分の方向性特徴を適用する．

1　はじめに

近年のパターン認識において，機械学習を用いた様々な手法が注目されている．機械学習は，文字通り機械に学習能力を付加するものである．パターン認識の分野では，その学習能力をカテゴリを分類する識別子として扱う．筆者らは，パターン認識における機械学習の中でも認識精度の高い手法の 1 つとされるサポートベクトルマシン [1] に着目し，同等の処理を類似度行列を用いた手法で行なえるよう検討する．

2　特徴抽出

特徴抽出には様々な手法が存在するが，本論文で用いたのは 16 成分の方向性特徴抽出 [2] である．文字認識における方向性特徴とは，文字のストローク成分にそったベクトル要素であることが多いが，本論文における方向性特徴は，文字パターンから背景パターンへ向かった法線ベクトルを表す．

ここで，特徴抽出の流れを示す．まず文字画像を平行・垂直方向に走査し，4 方向の方向性特徴成分を抽出する (図 1 の 1,2,4,8 に相当)．次に，4 方向のベクトル成分を組み合わせ，斜め方向を含めた 8 方向のベクトル成分へと分類する．図 1 と表 1 に，線素接続パターンおよびそれに対する方向成分への対応表を示す．図 1 の正方形は，任意の画素と隣接画素を表す．それぞれ，中央は対象画素，黒が文字画素，白が背景画素を示すものとする．

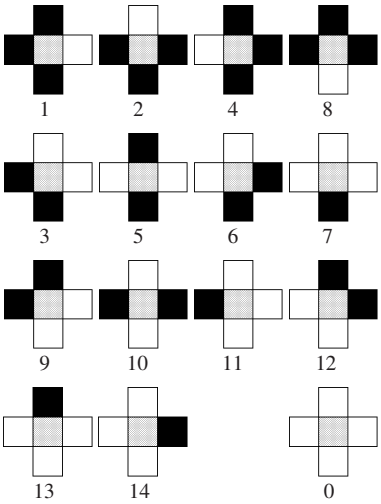
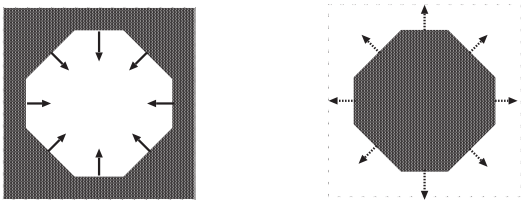


図 1：注目点における局所パターンと方向成分の関係

表 1：割り当て数値と方向成分の対応表

	→	↗	↑	↖	←	↙	↓	↘
0								
1	○							
2			○					
3		○						
4					○			
5	○				○			
6				○				
7		○		○				
8							○	
9								○
10			○				○	
11		○						○
12						○		
13						○		○
14				○		○		



(a) 相対成分有り　　(b) 相対成分無し

図 2：相対成分と方向性特徴

先の処理で 8 つに分類された特徴成分を，相対する方向成分の有無により更に 2 つに分類する．相対成分とは，自身と逆のベクトル要素をもつ方向性特徴である．図 2 は，相対成分の有無を表したものである．8 方向の特徴成分を矢印で表記し，相対成分がないものに関しては破線を用いている．16 方向特徴成分の値は，相対するベクトル同士の距離となる．相対するベクトルをもたない特徴成分は 0 とおく．よって，相対成分をもたない特徴成分は実際には使用されない．

3　学習サンプルの選択

本論文では，学習サンプル集合から各サンプルの類似度を用いて標準パターンに使用するものを分類する手法を提示する．

サポートベクトルマシンでは，2 種類のデータ集合から識別子 (サポートベクトル) を構成し，それを用いて分別するものであるのに対し，本論文で提示する手法では複数のデータ集合

から識別子を構成している．ここで，本論文で用いる識別子を識別参照パターンと称す．識別参照パターンを選択するためには，学習サンプル間の類似度を使用する．算出した類似度が，任意に設定する閾値より大きいサンプルを識別参照パターンの候補として選出する．

図 3 は，類似度の算出をニューロンモデルを用いて表したものである．ここで，図 3 および 4 の $x_\circ$ は学習サンプル，x_* を識別参照パターンを表すものとする．$||x_*|| = 1$ とした場合，$x_\circ$ と x_* の内積は 1 を最大値とする類似度 $S_{\circ,*}$ となる．閾値を h とすると，類似度が $1-h$ の範囲であるサンプルは識別参照パターンの候補とされる．

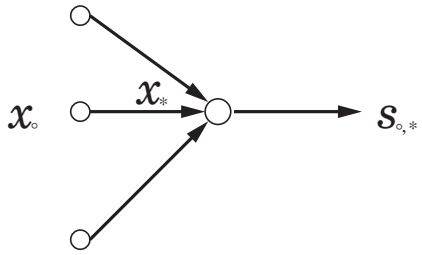


図 3：ニューロンモデル

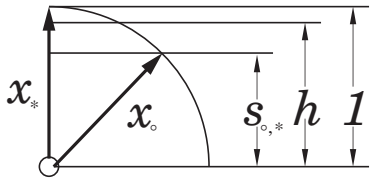


図 4：類似度モデル

次に，本手法の主な流れを式を用いて示す．先も述べたように，本論文では識別参照パターンの選択に類似度をを用いた手法をとっている．学習サンプルの類似度を関数 S とするとき，以下の式で表すことができる．

$$S_{(cn,i,cn2,j)} = \frac{\sum x_{cn,i} \cdot x_{cn2,j}}{||x_{cn,i}|| \cdot ||x_{cn2,j}||} \tag{1}$$

$cn \cdot cn2$ はカテゴリ番号，$i \cdot j$ はサンプル番号を表す．

任意のカテゴリ cn における対象サンプルとそれ以外カテゴリ $cn2$ との類似度 S の中で，類似度が最大となるサンプル番号 j を検出し，その類似度を関数 P として表す．

$$P_{(cn2,j)} = \max\{S_{(cn,i,cn2,j)} |\ cn2 \neq cn\ \} \tag{2}$$

他カテゴリとの最大類似度 P は，識別参照パターンを選出するための閾値として使用する．類似度 S が $c1$ より大きく，かつ閾値 $P_{(cn2,j)} + c2$ よりも大きい (対照とするサンプルの類似度が，他カテゴリの最良類似度値 $+c2\%$ 以上) 場合のサンプルを識別参照パターンの候補とし，候補リスト行列 Y に 1 を出力する．候補でないものには 0 を出力する．$c1$，$c2$ は，任意に決定するパラメータである．式 (3) の I,J は，ともにサンプルの最大数を表す．

$$Y = \begin{pmatrix} y_{11} & y_{21} & \cdots & y_{J1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1I} & y_{2I} & \cdots & y_{JI} \end{pmatrix} \tag{3}$$

$$Y = \begin{cases} 1 & ((S > c1) \ \wedge \ (S > P_{(cn2,j)} + c2)) \\ 0 & (otherwise) \end{cases} \tag{4}$$

図 5，図 6 は，各サンプルの組み合わせ毎の識別参照パターン候補の有無を表したものである．各パラメータは，$c1 = 0.85$ (最大類似度を 1 とした場合)，$c2 = 0$ となっている．使用データは，電総研手書き文字データベース **ETL6** の手書き数字部分 (0 ～ 9) で，偶数番目のサンプルを各カテゴリ毎に 70 文字 (計 700 文字) を採用している．図 5 にある横軸 · 縦軸の数値は，各カテゴリ値を表す．各点はカテゴリ毎のサンプルの情報であり，赤は $Y_{ji} = 0$ の場合，緑は $Y_{ji} = 1$ の場合，桃色は他カテゴリとの組み合わせであり，識別参照パターンの候補には含まれないことを表している．また，黒のラインは同一データであることを表す．図 6 は，図 5 と同条件でカテゴリが 0 の場合のみを表示したものである．ただし $Y_{ji} = 1$ の場合は，類似度の高低を緑の濃淡で表現している．

先の処理で得た候補リスト Y_{ji} を用いて，識別参照パターンを選出する．図 7 に，簡単のため 6×6 の大きさで表した行列 Y_{ji} のモデルを示す．注目サンプルごとに $Y_{ji} = 1$ となる要素を検出し，その数が最も多いサンプルを第一の識別参照パターンとする (図 7(1))． j と i は，これまでと同様に学習サンプル番号を表し，添字の ji は同一カテゴリ間の各サンプルの組み合わせを表す．また，図 7 の色付きの要素は検出可能箇所を表す．二つめの以降の識別参照パターンの選出は，これまでに選出された識別参照パターンを元に行なわれる．一つ前に検出したサンプル (識別参照パターン) について，$Y_{ji} = 0$ となる場合の組み合わせを抽出し (図 7(2))，検出対象サンプルとする．すでに選んだサンプル以外の中から，前回と同様に注目サンプルごとに $Y_{ji} = 1$ となる要素が最大になるサンプルを検出し，新たな識別参照パターンとする (図 7(3))．この処理を検出される Y_{ji} の要素が全て 0 になるまで繰りかえす．識別参照パターンは，各カテゴリごとに繰りかえしの回数分選出されることになる．一度検出の対象外になったサンプルは，以降の検出処理には含まれない．

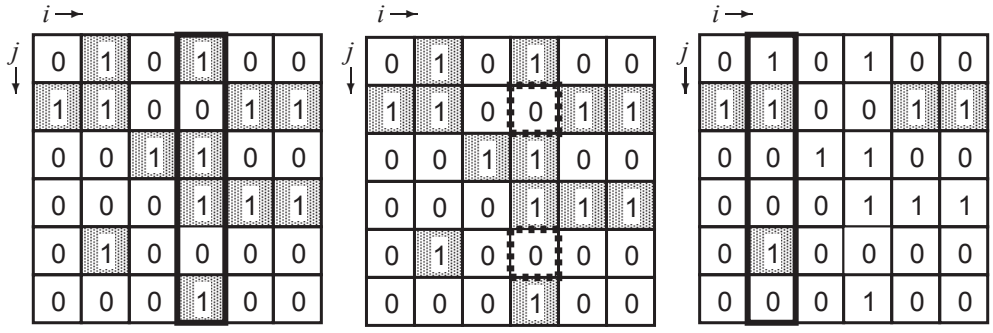


図 7：識別参照パターン選出モデル

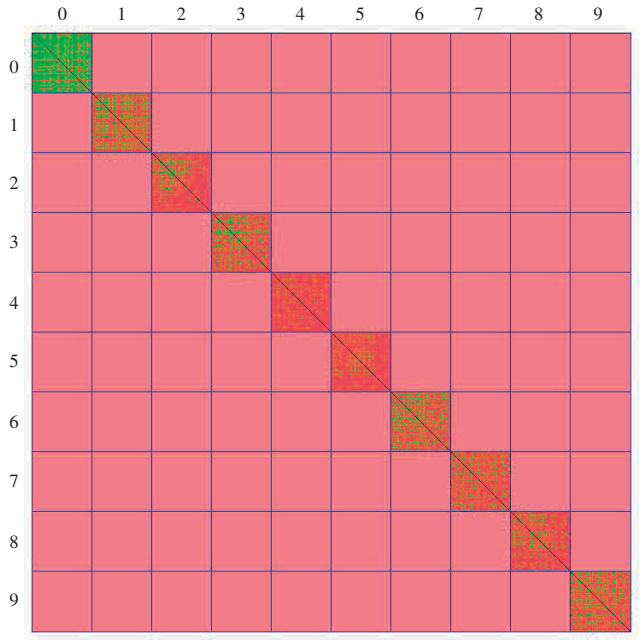


図 5：各カテゴリ毎の識別参照パターン候補

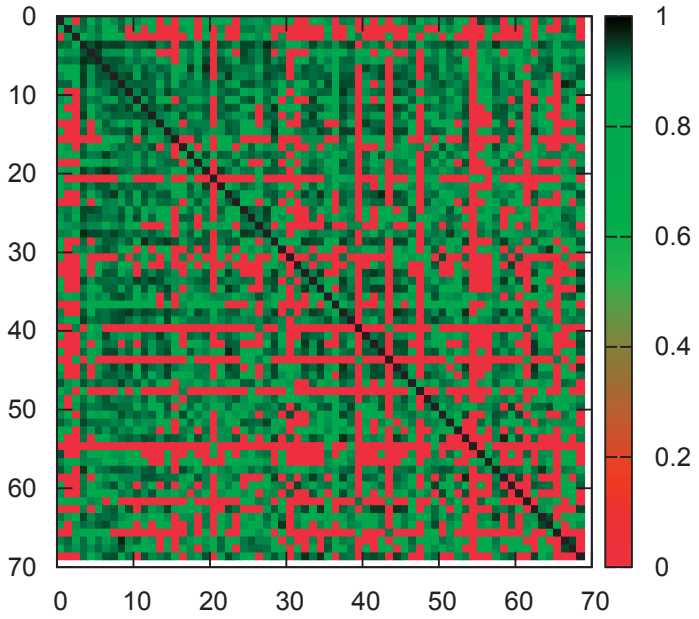


図 6：カテゴリ 0 における識別参照パターン候補

4　シミュレーション実験

実験に使用したデータは、電総研手書き文字データベース **ETL6** の手書き数字部分 (0 ～ 9 までの 10 カテゴリ) である．データ数は 1 カテゴリあたり 1383 文字あり，偶数番目と奇数番目の 2 種類のデータに分けて使用した．２種類のデータは各 691 文字とし，1383 文字目のデータは除くものとする．つまり学習サンプルは，1 つ当たり 6910 文字 (10×691 文字) で構成されることになる．(以後，偶数番目のデータを e，奇数番目のデータを o と表記) 元データサイズは 64×63 画素，実験では文字範囲の切り出し処理をしたのち，8×8 画素の範囲 (作業領域 14×14 画素の中央に配置する) に正規化する．

その他の処理として，類似度の算出のさいに 3×3 画素の範囲でガウシアンフィルタを適用した (ぼけ処理との組み合わせ [3])．実験に使用した重み係数を表 2 に示す．$|h|$ は水平方向，$|v|$ は垂直方向に向かう注目点からの距離とする．なお，本実験では摂動相関法 [4] との組み合わせは行なっていない．

実験では，パラメータ別に標準パターンを作成し，認識シミュレーションを行なった．パラメータとは，式 (4) における $c1$，$c2$ を表す．本論文ではパラメータ毎に認識実験を行ない，それぞれを実験 A，B，C，D と称す．詳細を表 3 に示す．ただし，最大類似度を 1 としたときの値とする．また，実験に使用した環境を表 4 に示す．

表 2：重み係数

$\|v\|$ \ $\|h\|$	0	1	2
0	15	12	6
1	12	9	4
2	6	4	2

表 3：各パラメータの値と実験データ

	実験 A	実験 B	実験 C	実験 D
$c1$	0.80	0.85	0.88	0.90
$c2$	0.05	0.05	0.05	0.05

表 4：実験環境

MPU	Intel(R) Xeon(R) CPU 5160@ 3.00GHz
内部記憶装置容量	2GB
コンパイラ	Intel(R) C++ Compiler Version 10.0

5　実験結果

表 5 と図 8 に学習サンプルの選択数を示す．項目は，識別参照パターンを用いた各実験とサポートベクトルマシン (以後，svm と表記) を用いたものを示す．サポートベクトルによる結果は，文献 [5] のライブラリを使用したものである．図 9 に選択された学習サンプルの例 (実験 D，偶数パターン) を示す．表示データは，正規化済みのものである．

$c1$ は，識別参照パターンの候補を選出するためのパラメータである．同一カテゴリ間の類似度が $c1$ よりも大きい場合は候補に選出される．その候補の中で最も相関の高いものが第一

表 5：学習サンプルの選択数 (a)

	実験 A	実験 B	実験 C	実験 D	svm
o	221文字	359文字	592文字	959文字	744文字
e	198文字	316文字	539文字	887文字	813文字

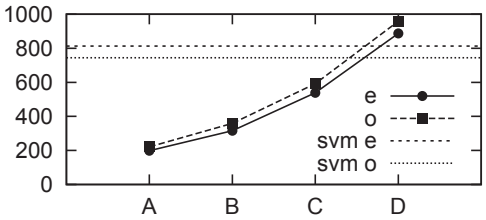


図 8：学習サンプルの選択数 (b)

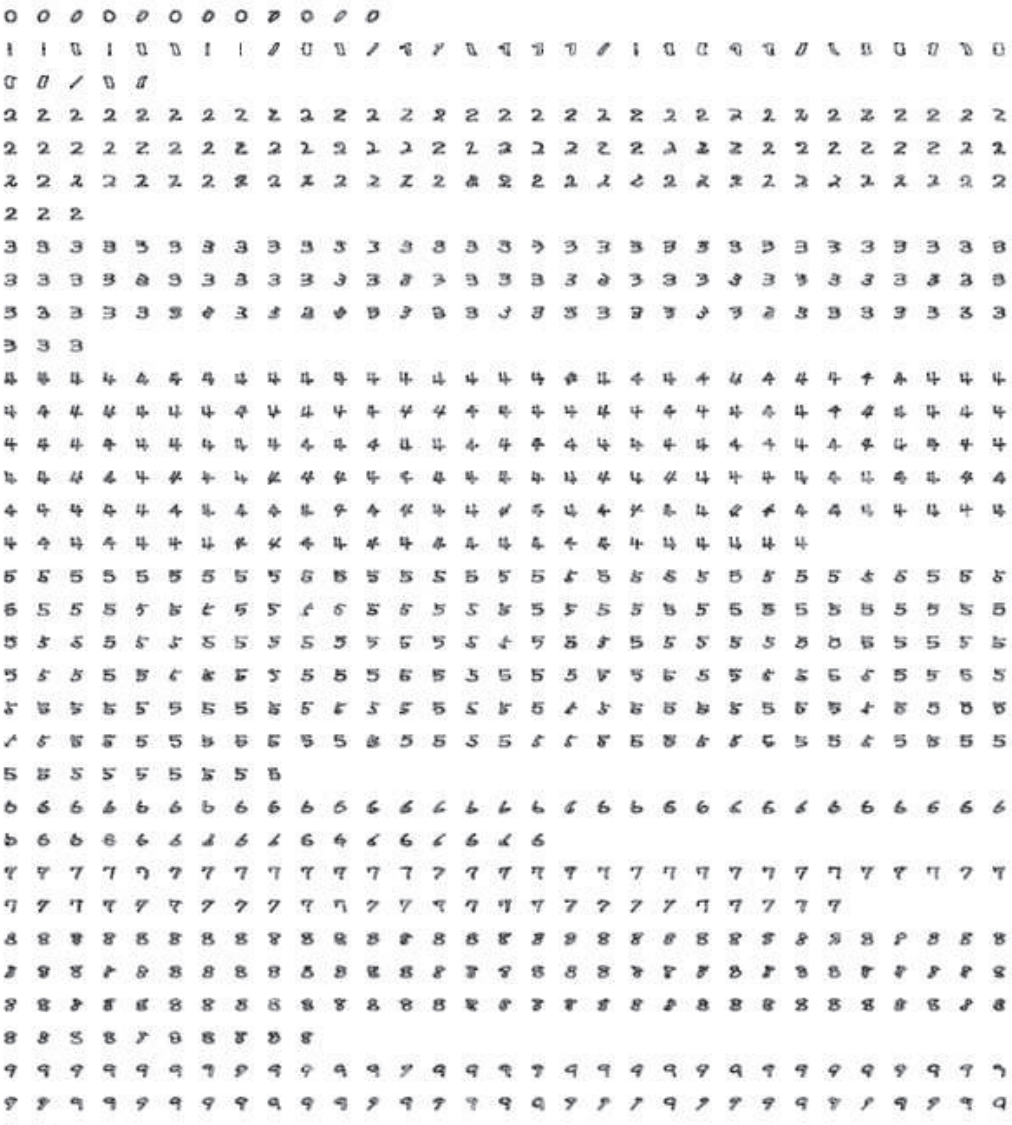


図 9：学習サンプルの選択例 (実験 D，偶数パターン)

識別参照パターンとして選ばれることになるが，それ以降は第一識別参照パターンにおいて相関の低かった範囲で再選択が行なわれる．よって，$c1$ の値が大きいほど第一の識別参照パターンの選出の範囲が少なく，それ以降の選出量が大きくなることになる．表 5 および図 8 から，パラメータ $c1$ の値が大きくなるほど学習サンプルの選択数も増大の傾向がみられることから，想定道理の結果が得られたといえる．

データの種類から比較すると，奇数データの方が数十文字程多く検出されている．その差は最大で実験 D の 72 文字であるが，この数値は全サンプル数の 1% 程度であり，データの違いによって類似度等も変化することから，この程度の差異は許容範囲であるといえる．また，svm による結果は実験 C と D の中間程度の選択数であった．

表 6 に学習サンプルの選択に要した処理時間を記す．何れの実験も 2,230 秒 (約 37 分) 程度と，選択される文字数による違いは見られない．

表 6：学習サンプル選択の処理時間

実験 A	実験 B	実験 C	実験 D
2,235.20s	2,235.15s	2,234.51s	2,237.15s

表 7 に，実験 A~D から生成した標準パターンを用いて認識シミュレーションを行なった結果 (誤読文字数) を示す．同時に，svm と従来法による認識結果も示す．従来法は前論文 [2] の内容のもので，本論文と同様に 16 方向の線素特徴を用いたものである．項目にある oo，oe，eo，ee は，認識における標準パターンと未知パターンの組み合わせを意味するもので，左側が標準パターン，右側が未知パターンを表す．

表 7：認識結果 (誤読文字数)

	実験 A	実験 B	実験 C	実験 D	svm	従来法
oo	1文字	3文字	0文字	0文字	0文字	26文字
oe	13文字	12文字	10文字	7文字	7文字	24文字
eo	15文字	9文字	9文字	7文字	8文字	26文字
ee	1文字	0文字	0文字	0文字	0文字	27文字

同一パターンの組み合わせは，実験 A,B では 1 文字前後の誤認識，実験 C,D においては誤認識なしという認識結果となった．複数のサンプルを標準パターンに適用することで，全体の平均化による標準パターンモデルではカバーすることの出来なかった文字まで対応可能となった．特に同一パターンの組み合わせについては，同一文字間の相関をとる場合が生じるため，その影響が顕著に表われることが確認できる．異なるパターンとの組み合わせにおいても，従来法と比べ誤読文字数の減少が確認できる．

実験 A~D の結果から，パラメータ $c1$ の値を高くすることで誤読文字に減少が確認できる．また，実験 D では svm と同等の結果が得られた．

表 8 および図 10 に，各実験の認識処理時間を示す．当然のことではあるが，学習パターンの選択数に比例して認識処理に要する時間も増加している．表 7 から，学習パターンの選択数を増やすことで正読率の改善が確認されたが，それは同時に処理時間が増大することを意味する．このことから，認識と処理時間の双方からみた最適な識別参照パターンを採択することが重要であると考えられる．

表 8：認識処理時間 (a)

実験 A	実験 B	実験 C	実験 D
4.04*s*	4.91*s*	6.23*s*	9.05*s*

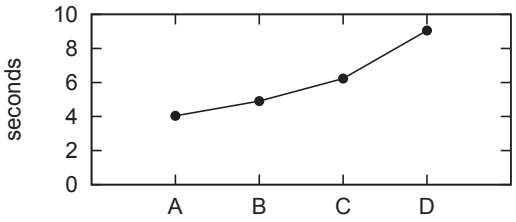


図 10：認識処理時間 (b)

図 9 は，実験 D において認識時に算出された標準パターンと未知パターンとの類似度をカウントしたものである．このデータは，正解カテゴリとの類似度から算出したものである．図 9 における最大類似度は，100 となっている．

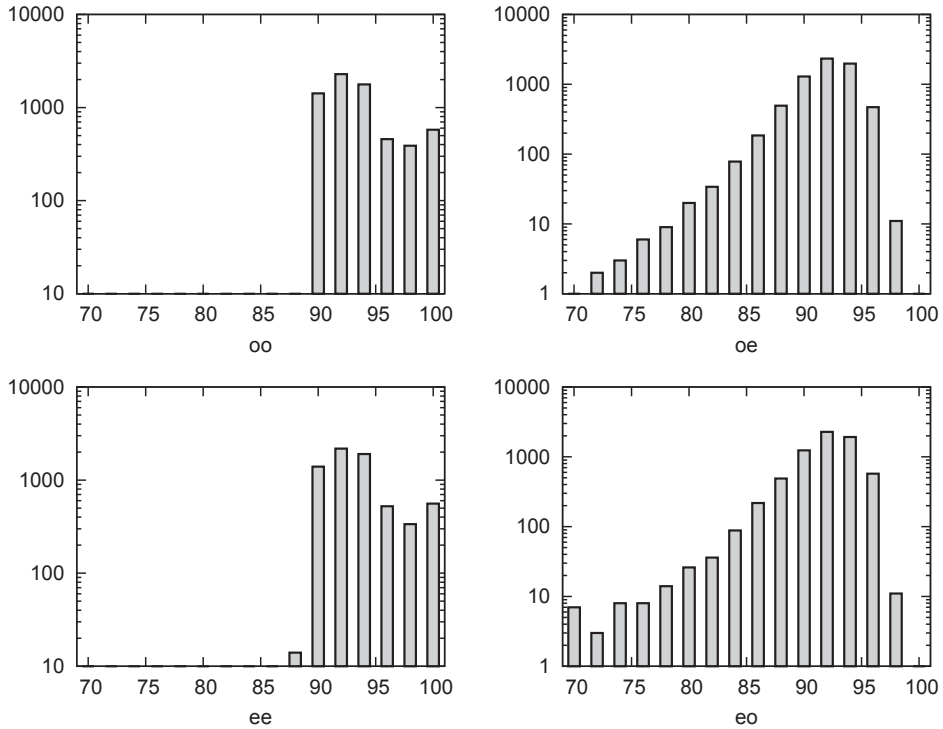


図 9：認識の類似度傾向 (実験 D)

同一パターンの組み合わせによる結果は，右寄り (類似度がほぼ 90% 以上) の傾向にあり，90 ～ 95% の範囲の組み合わせにピークが確認できる．先にも述べたように，同一パターン同士の組み合わせは，同一文字間の相関をとる場合が生じる．同一文字の組み合わせの場合，類似度は 100% となる．つまり，グラフの最右端は同一文字となったサンプル数であるといえる．一方，異なるパターンとの組み合わせの結果は，同一文字が存在しないため類似度が 100% のサンプルは存在せず，同一パターンと同じ 90 ～ 95% の範囲をピークにして 70% までゆるやかに減少している．

6　まとめ

機械学習の考え方を元に，類似度行列を用いて学習サンプルを採択する手法を検討した．シミュレーション実験の結果，他パターンの組み合わせにおいても誤読文字数 7 文字と，サポートベクトルマシンと同等の結果が得られた．また処理時間についても，対判定であるサポートベクトルマシンに比べて処理が簡潔であり，効果があったと考えられる．

今後の課題として，他手法とのさらなる比較，摂動相関法 [4] との組み合わせや正規化範囲の変更等が挙げられる．また，手書き数字以外のデータを用いた実験，**ETL6** 以外のデータを用いた実験ついても検討したい．

参考文献

[1] 栗田：” サポートベクターマシン入門”，
http://home.hiroshima-u.ac.jp/tkurita/lecture/svm/index.html

[2] 二階堂，北村，中島，安田：” 線素特徴の拡張とその効果”，明星大学研究紀要 —情報学部—第 19 号

[3] 安田，藤沢，” 文字認識のための相関法の一改良”，信学論，62-D，3，pp217-224，Mar. 1979．

[4] 北村，二階堂，中島，安田：” 文字認識のための適応摂動相関法の提案”，明星大学研究紀要 —情報学部—第 14 号

[5] 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://csie.ntu.edu.tw/˜cjlin/libsvm