

Title	Support Vector Machine に関する時系列データの認識に関する研究
Author(s)	中井, 浩一
Citation	
Issue Date	2000-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1334
Rights	
Description	Supervisor: 下平 博, 情報科学研究科, 修士

修士論文

Support Vector Machine による
時系列データの認識に関する研究

指導教官 下平博 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

中井浩一

2000年3月修了

要旨

本論文では、パターン認識の分野において音声や文字などの時系列パターンを扱う新しい識別器を提案する。現在、固定長データに対して高い識別性能を示す識別器として Vapnik らによって提案された Support Vectors Machine(SVM) がある、この SVM に注目し時系列パターンの認識に用いるモデルとして SVM 連鎖モデルを提案する。そして、認識実験を通して提案手法の性能を評価する。

目次

1	はじめに	1
1.1	この論文の構成	2
2	Support Vector Classifier (SVC)	3
2.1	線形 SVC	3
2.2	非線形 SVC	6
2.3	構成的誤り最小化理論 (SRM 理論)	8
3	時系列パターンに対応した SVC	10
3.1	概要	10
3.2	SVC 連鎖モデル	10
3.2.1	クラス帰属度	12
3.2.2	遷移の重み	14
3.3	SVC 連鎖モデルの認識アルゴリズム	14
3.4	複数カテゴリの認識 (3 カテゴリ以上)	14
3.5	SVC 連鎖モデルの学習アルゴリズム	16
3.5.1	Viterbi アルゴリズム	16
3.5.2	CSVC 学習アルゴリズム	16
3.6	学習アルゴリズムの収束性	17
4	手書き数字文字の認識実験	19
4.1	2次元特徴による評価実験	21
4.1.1	重み λ に関する検討	21
4.1.2	カーネル関数の次数 d に関する検討	22
4.1.3	反復学習の回数に関する検討	23
4.1.4	HMM との性能比較	24

4.2	4次元特徴による評価実験	25
4.2.1	重み λ に関する検討	25
4.2.2	カーネル関数のパラメータ d に関する検討	25
4.2.3	遷移の重みに関する検討	27
4.2.4	クラス帰属度関数に関する検討	28
4.2.5	誤り項の正規化法に関する検討	29
4.2.6	HMM との性能比較	30
5	実験の考察	31
5.1	認識率に関する考察	31
5.2	CSVC の収束性に関する考察	31
5.3	学習アルゴリズムに関する検討	34
5.3.1	学習アルゴリズムの問題点	34
5.4	学習アルゴリズムの改善	38
5.5	モデルの評価法に関する考察	39
5.5.1	実験 1	40
5.5.2	実験 2	42
5.5.3	実験 3	46
6	結論	49
6.1	研究結果	49
6.2	今後の課題	50

第 1 章

はじめに

Support Vector Machine[1, 2, 4, 3] は 1995 年に Vapnik らによって提案された新しいパターン認識の手法であり、2 クラスのパターン識別問題において高い認識性能を示すことから近年注目をされている。SVM を利用した 1 つの手法として回帰推定を行う Support Vector Regression (SVR)[11] があり、これと区別をするために識別を行う SVM は Support Vector Classifier (SVC) と呼ばれることがある。以下、本論文においても SVM は総称の意味で用い、識別器を指す場合は SVC と記すことにする。

SVC は、学習が 2 次の凸型計画問題として定式化されているため局所解に落ちいることなく大域的最適解を求めることができる点や、構成的リスク最小化 (Structural Risk Minimization: SRM) 原理に基づいた学習を行なうために汎化能力が高い点に特徴がある。

SVC は文字認識や画像認識等の様々な応用分野において、従来法を上回る高い識別性能を示すことが確認されつつある [5]。しかし、SVC は特徴量ベクトルの次元数が固定であることを仮定している。そのため、音声信号やオンライン手書き文字のように特徴量がベクトルの時系列として与えられ、系列の長さが可変であるようなデータを扱うためには、何らかの対処が必要である。SVC と同様に固定長データのみを扱う識別器を時系列データの認識に応用した例としては、ニューラルネットワークによる音声認識などがある [6, 7]。SVC を音声認識に応用した例としては、[8, 9] があるがこの研究では時系列データである音声信号を時間軸に線形正規化することで固定次元のベクトルに変換して音素認識を試みている。そのため可変長データを固定長データに変換することによる情報の欠落などの検討が必要になる。

そこで本研究では時系列信号を直接扱える SVC の枠組として音声認識の分野で優れた認識性能を示している隠れマルコフモデル (hidden Markov Model: HMM)[16, 17] と同様に、時系列信号を複数の定常信号源の連鎖として扱うモデルの開発を試みる。開発する時

系列信号を扱う SVC のモデルは従来の固定長データを扱う SVC をそのまま利用するため固定長データを認識する SVC が従来法を上回る性能を示した場合と同様に時系列データについても従来法よりも優れた認識性能を示す可能性がある。

本研究では、時系列データを扱う SVC のモデルとして SVC 連鎖モデル (Chain SVC: CSVC) を提案し、モデルのアルゴリズム及び定式化を行ない、手書き数字の認識実験を通して提案モデルの有効性を示すことを目的とする。

1.1 この論文の構成

本論文では、第 2 章では従来の固定長データを扱う SVC の概念とアルゴリズムについて述べる。第 3 章では時系列データを識別する新しい SVC のモデルとして複数の SVC を接続して構成した SVC 連鎖モデル (Chain SVC: CSVC) を提案し、SVC の識別関数を評価関数として Viterbi アルゴリズムと同様の認識 / 学習のアルゴリズムと定式化について説明を行なう。第 4 章では、時系列データの認識実験として手書き数字の認識実験を行い CSVC と HMM の比較を行ない CSVC の評価と考察を行う。第 5 章では、実験結果から CSVC の認識性能、及び学習アルゴリズムの検証を行なう。第 6 章では本稿全体の考察を行ない、今後の解題について述べる。

第 2 章

Support Vector Classifier (SVC)

本章では固定長データを対象にした従来の SVC について説明をする。Vapnik 等によって 1995 年に提案された Support Vector Classifier (SVC) は新しい統計的学習理論に基づくパターン認識手法として近年注目を集めている。SVM は、large-margin classifier の一種で、2 クラスのパターン識別問題においてパターン空間上に非線形な識別境界面を構成するノンパラメトリックな学習法である SVC の特徴として、

- 学習問題の定式化が構成的リスク最小化 (Structural Risk Minimization: SRM) 原理 [10] に基づいており、学習資料に対する認識誤り (経験的誤り) と汎化性能 (VC 信頼) の両面から学習の最適化が行われるため、経験的誤りについてのみ学習を行なう手法に比べ汎化性能の点で優れている。
- 学習が 2 次の凸計画問題として定式化されているため、大域的最適解を比較的容易に求めることができる。
- 線形分離不可能なパターンを高次元空間に非線形写像し写像した先の高次元空間で線形識別を行なう、また膨大な演算を伴う非線形写像を、それと等価な演算を高速に行なうカーネル関数により実現している。

が挙げられる。

2.1 線形 SVC

最初に、パターンの非線形写像を行わずに元のパターン空間上で線形識別を行なう SVC について説明をする。

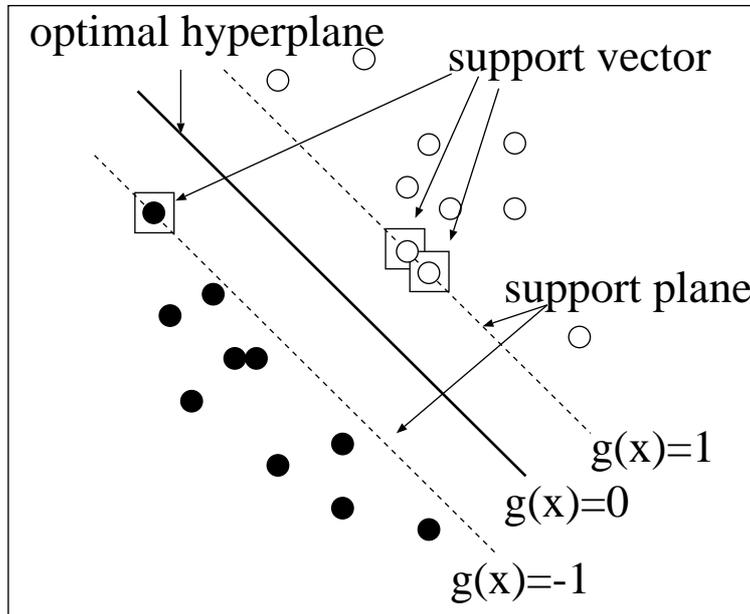


図 2.1: support plane と optimal hyperplane

SVC は入力されたパターンをサポートプレーン (support plane) と呼ばれる 2 つの超平面を構築し、2 クラスの境界となる最適分離超平面 (optimal hyperplane) を 2 つのサポートプレーンの中間に決定する (図 (2.1))

SVC にラベル付けされた N 個の学習資料 (x_i, y_i) $i = 1, \dots, N$ が与えられたとする、ここで $x_i \in R^d$ は i 番目の d 次元パターンベクトルであり $y_i \in \{-1, 1\}$ は i 番目のパターンベクトルのクラスラベルでありどちらのクラスに属するパターンベクトルであるかを表している。

パターン空間上の超平面は、式 (2.1) で与えられる SVC の識別関数で表現できる。

$$g(x) = w \cdot x + b \quad (2.1)$$

ここで“ \cdot ” はベクトルの内積を意味し、 w は超平面の法線ベクトルを表し b はスカラーの定数である。 w 、 b は識別関数のパラメータであり SVC は学習によりこれらのパラメータを決定する。与えられた学習資料が線形分離可能ならば学習資料は、 $w \cdot x + b = +1$ と $w \cdot x + b = -1$ となる 2 つの超平面 (サポートプレーン) により、それぞれ的一方の側に該当するクラスの資料が分布する様に分離することができる。SVC はこのような分離が可

能な超平面の中から2つのサポートプレーン間の距離(クラス間距離)を最大にするパラメータを決定する、またサポートプレーン上にある学習資料をサポートベクター (support vector) と呼ぶ。次に学習データが線形分離不可能な場合は式(2.1)のような線形識別関数で誤分類なく識別することは不可能である、そこで誤分類したパターンの support plane からの逸脱量を表す変数として $\xi_i \geq 0, i = 1, \dots, N$ を導入し以下のような不等式による分離を行なう。

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq 1 - \xi_i \quad (\text{if } y_i = +1) \quad (2.2)$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 + \xi_i \quad (\text{if } y_i = -1) \quad (2.3)$$

この場合 SVC は、リスクの上限の最小化のため式(2.4)の最小化を行なうように学習する。学習の定式化は次式のような2次の凸型計画法となるため最適解を求めることが可能である。

$$\text{Minimize} \quad L = \frac{\lambda}{2} \mathbf{w} \cdot \mathbf{w} + (1 - \lambda) \sum_{i=1}^N \xi_i \quad (2.4)$$

$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad (2.5)$$

$$\xi_i \geq 0, \quad \text{for } i = 1, \dots, N \quad (2.6)$$

ここで式(2.4)の第1項はサポートプレーン間距離 $\frac{2}{\sqrt{\mathbf{w} \cdot \mathbf{w}}}$ を2乗し逆数をとったものであり、第2項は学習誤りの総和である、 λ は $[0, 1]$ の範囲をとる係数で第1項と第2項の重みを表している。

この最適化問題は Lagrange の未定乗数 (α) により以下のように書き換えることができる。

$$L = \frac{\lambda}{2} \mathbf{w} \cdot \mathbf{w} + (1 - \lambda) \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i \{y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 + \xi_i\} - \sum_{i=1}^N \beta_i \xi_i \quad (2.7)$$

また鞍点では、 \mathbf{w}, b, ξ は以下の状態を満たす。

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \lambda \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0 \\ \mathbf{w} &= \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i = 0 \end{aligned} \quad (2.8)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^N \alpha_i y_i = 0 \quad (2.9)$$

$$\frac{\partial L}{\partial \xi_i} = (1 - \lambda) - \alpha_i - \beta_i = 0 \quad (2.10)$$

これにより式 (2.7) は次式のようになる

$$\text{Maximize } L = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \quad (2.11)$$

$$\text{subject to } \sum_{i=1}^N \alpha_i y_i \quad (2.12)$$

$$0 < \alpha_i < \frac{1 - \lambda}{\lambda}, \quad i = 1, \dots, N \quad (2.13)$$

この最適化問題を解くことにより最適超平面を与える α が求まる、ここでサポートベクター以外の α の値は0となるため式 (2.8) は

$$\mathbf{w} = \frac{1}{\lambda} \sum_{i=1}^{N_{sv}} \alpha_i y_i \mathbf{x}_i \quad N_{sv}: \text{number of support vector} \quad (2.14)$$

となる、さらに定数 b は次式で与えられる。

$$b = \frac{1}{2\lambda} [(\mathbf{w} \cdot \mathbf{x}_{sv}(+1)) + (\mathbf{w} \cdot \mathbf{x}_{sv}(-1))] \quad (2.15)$$

ここで、 $\mathbf{x}_{sv}(+1)$ はクラスラベル+1のクラスのサポートベクターを表し $\mathbf{x}_{sv}(-1)$ はクラスラベル-1のクラスのサポートベクターを表している。

求められた \mathbf{w} と b を用いて SVC のクラス分類は次式により行なわれる。

$$\text{sgn}\left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b\right) \quad (2.16)$$

2.2 非線形 SVC

以下、第??節では従来の SVC の原理について概説した後、時系列データを認識する CSVC(3節)、手書き数字文字データを使った評価実験結果 (4節) について述べる。

前節では線形識別関数を用いた SVC の説明をしたが、SVC は線形識別関数以外の識別関数を関数形を仮定することなく構成することができる。本節ではカーネル関数による非線形写像を用いた非線形 SVC について説明する。

非線形 SVC では、非線形写像関数 ($\Phi: \mathcal{R}^n \rightarrow \mathcal{R}^{n'} (n \ll n')$) によって、データ \mathbf{x} を高次元空間上に写像する、このような写像により元の空間では線形分離不可能な分布を写像先の高次元空間では線形分離可能な分布にすることができるため、高次元空間での線形識別関数による識別が可能になる。この時、構成された超平面による境界面は高次元空間では線形であるが元の次元の空間では非線形な境界面となっている (図 2.2)。しかし、実際にこのような写像には膨大な量の演算が必要になるので実現するのは非常に困難にな

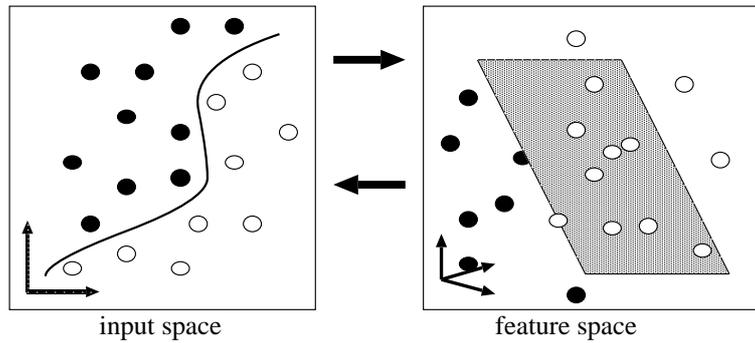


図 2.2:

る。このような問題に対して SVC は $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ となる性質を有するカーネル関数を利用することにより、もとのベクトル空間 (\mathcal{R}^n) 上で等価で高速な処理を実現している [?]。SVC で用いられるカーネル関数の具体的な例として以下の 3 つの関数を紹介する。

- polynomial function

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d \quad (2.17)$$

- radial basis function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{|\mathbf{x}_i - \mathbf{x}_j|}{2\gamma^2}\right) \quad (2.18)$$

- two layer neural nets

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa \mathbf{x}_i \cdot \mathbf{x}_j - \delta) \quad (2.19)$$

なお、非線形 SVC の定式化は線形 SVC の内積計算がカーネル関数による計算に置換されるだけで線形 SVC の場合と同型であるので、最適超平面を求める最適化問題は式 (2.4) から次式のようになる。

$$\text{Minimize} \quad L = \frac{\lambda}{2}K(\mathbf{x}_i, \mathbf{x}_j) + (1 - \lambda) \sum_{i=1}^N \xi_i \quad (2.20)$$

$$\text{subject to} \quad y_i \left(\sum_{i=1}^N \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b \right) \geq 1 - \xi_i \quad (2.21)$$

$$\xi_i \geq 0, \quad \text{for } i = 1, \dots, N \quad (2.22)$$

そして識別関数は式 (2.1) から次式のようになる。

$$g(\mathbf{x}) = \sum_{i=1}^N \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}_j) + b \quad (2.23)$$

2.3 構成的誤り最小化理論 (SRM 理論)

先に述べた通り SVC は SRM 理論に基づいた学習を行なうために、汎化性能の点で優れている。学習の際に経験的誤りに対する最小化を行なう識別器の問題点として、汎化性能が低い (学習データに対する認識率は高いが評価データに対する認識率が低い) とゆう点が挙げられる。これは経験的誤りの最小化は誤りの期待値 (リスク) の最小化でないためである。これに対して N 個の学習資料から識別器のモデルパラメータ α が与えられた場合、Vapnik によりこの識別器のリスクの上限が次式のように経験的誤りと汎化性能の信頼区間の和であることが示された。

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(\log(2l/h)) - \log(\eta/4)}{N}} \quad (2.24)$$

式 (2.24) の右辺第 2 項は VC 信頼と呼ばれる項であり、 η は $0 \leq \eta \leq 1$ となる定数であり h は VC 次元と呼ばれる非負の整数である。式 (2.24) の第 1 項と第 2 項は h についてトレードオフの関係にあるため (図 2.3)、リスクの上限の最小値は、2 つの項の和の最小値を与える VC 次元を求める問題になる。そしてリスク上限の最小値を与える VC 次元を求めるのが SRM 理論である。

SVC の VC 次元については次式のような関係がある。

$$h \leq \min\{[D^2/M^2], d\} + 1 \quad (2.25)$$

ここで d はデータの次元数であり、 M はサポートプレーン間の距離である。つまり SVC の学習で用いた評価量は式 (2.24) に対応しており式 (2.4) のサポートプレーン間距離の

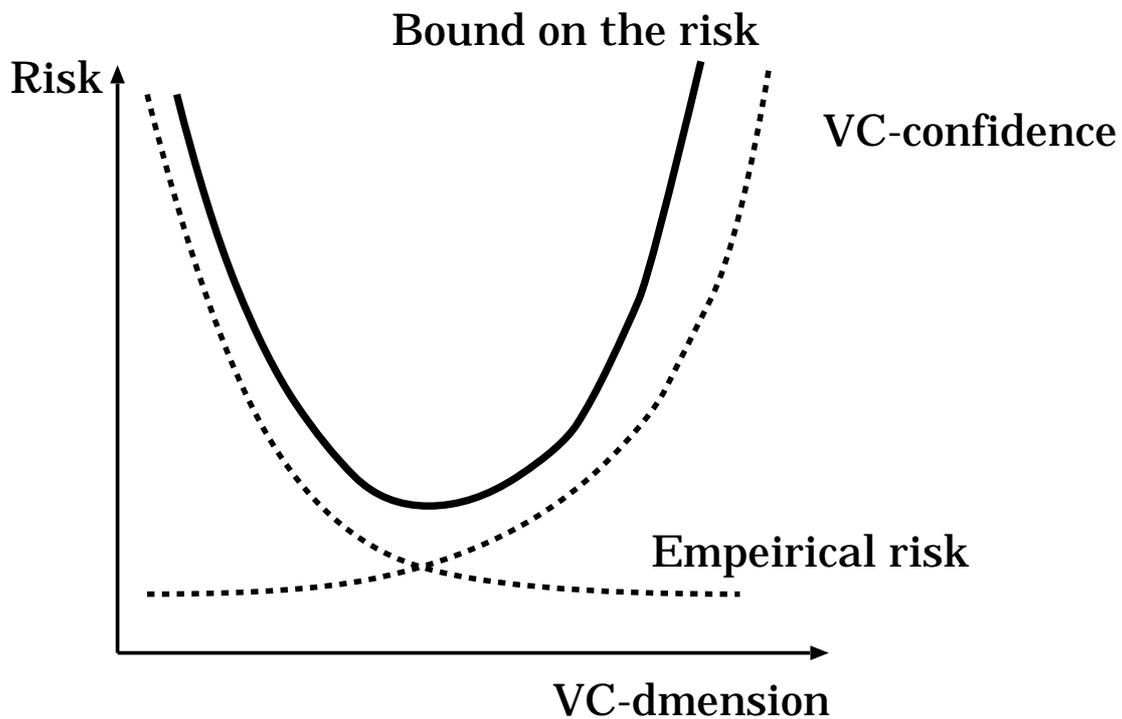


図 2.3: 経験的誤りと VC 信頼

経験的誤りと VC 信頼はトレードオフの関係にあり、
 VC 次元を大きくすると経験的誤りは減少するが VC 信頼が増化する、
 逆に VC 次元を小さくすると VC 信頼は減少するが経験的誤りが増化する。

逆数である第 1 項 $\frac{\lambda}{2} \mathbf{w} \cdot \mathbf{w}$ の最小化は VC 次元の上限の最小化を意味する、また第 2 項 $(1 - \lambda) \sum_i \xi_i$ が経験的誤りを意味している。このため SVC の学習は、SRM 理論に基づいた汎化能力と経験的誤りの 2 つの評価量の最適化を行うことになる。

第 3 章

時系列パターンに対応した SVC

3.1 概要

前章で述べた SVC は固定長データ (データの次元数一定) のみを対象とした識別器であり、時系列データを扱うことができなかった。SVC が時系列データを扱えない理由としては、従来の SVC が対象とするデータの次元数が一定である事を前提として成り立つ識別器であるのに対して、時系列データは系列長が変化するとゆう特徴から、系列長の変化が次元数の変化に相当するため SVC が時系列データを扱うことはできなかった。

本章では時系列データに対応した SVC のモデルとして提案する SVC 連鎖モデルについて説明する。

3.2 SVC 連鎖モデル

最初に時系列データについての考察を行なう。時系列信号を等時間間隔で標本化して得られる特徴量 (ベクトル) をフレーム特徴量とよび、その時系列を $x_t, t = 1, \dots, n$ と記すことにする。この場合、フレーム特徴量 x_i の次元数は時間によらず一定であるが、時間方向の長さ n はデータが生成される度に変わるので、ベクトルの時系列データ $X = (x_1, x_2, \dots, x_n)$ は可変長のベクトル系列となる。つまり時系列データはフレーム特徴量の系列であり、データが生成される度に変わるのはフレームの個数である。

このような時系列データを SVC を用いて認識するために、フレーム特徴量が固定長データであることに注目し、時系列データをフレーム特徴量単位で認識する手法を考える。この手法は認識対象がフレーム特徴量であるため従来の固定長データを扱う SVC が利用できる、そこで複数の SVC を時間的に切り替えることで時系列データの認識を行う SVC

連鎖モデル (Chain SVC: CSVC) を提案する。

ここでは手書き数字認識を具体例として CSVC の説明を行なう図 3.1。手書き数字認識では時系列データとしてフレーム特徴量を 2 次元速度ベクトルとしたベクトル系列を扱う、図中の小さな矢印は特徴量として速度ベクトル (位置ベクトルの時間差分) を方向と線分の長さで示しており、各文字について 3 個のデータ (data1 ~ data3) の速度ベクトル系列が表示されている。また、縦線は入力ベクトル系列に対して 3 個の SVC が識別を担当する区間の境界を示している。

図 3.1は、手書き数字 "2" と "3" の時系列パターンを識別するための 3 状態から成る CSVC の例を示している。各状態 s_i , $i = 1, 2, 3$ にはそれぞれ 1 つの SVC が割当てられており、時系列信号の性質から状態変化 (状態遷移) の順番は時間に関して非減少で、いわゆる left-to-right 型の構造となっている。CSVC はパラメータとして SVC の遷移の重みとフレーム特徴量のクラス帰属度の 2 種類を持つ。遷移の重みは遷移の容易さを表わす非負の変数が定義されており、 a_{ij} は状態遷移 $s_i \rightarrow s_j$ に関する重みを意味する。

図から分かるように、CSVC は隠れマルコフモデル (HMM) と同様の構造を有している。HMM との大きな相異点は、HMM が各状態に信号生成の出力確率密度関数を与えているのに対して、CSVC では各 SVC の識別対象になっているクラスから信号が生成される度合いを表すクラス帰属度関数を用いることである。

SVC 連鎖モデルを時系列パターンの認識に用いることの利点は、後述するように SVC の定式化を大幅に変更する必要が無いことその他、HMM で用いられている様々なアルゴリズムが適用できる可能性があることである。

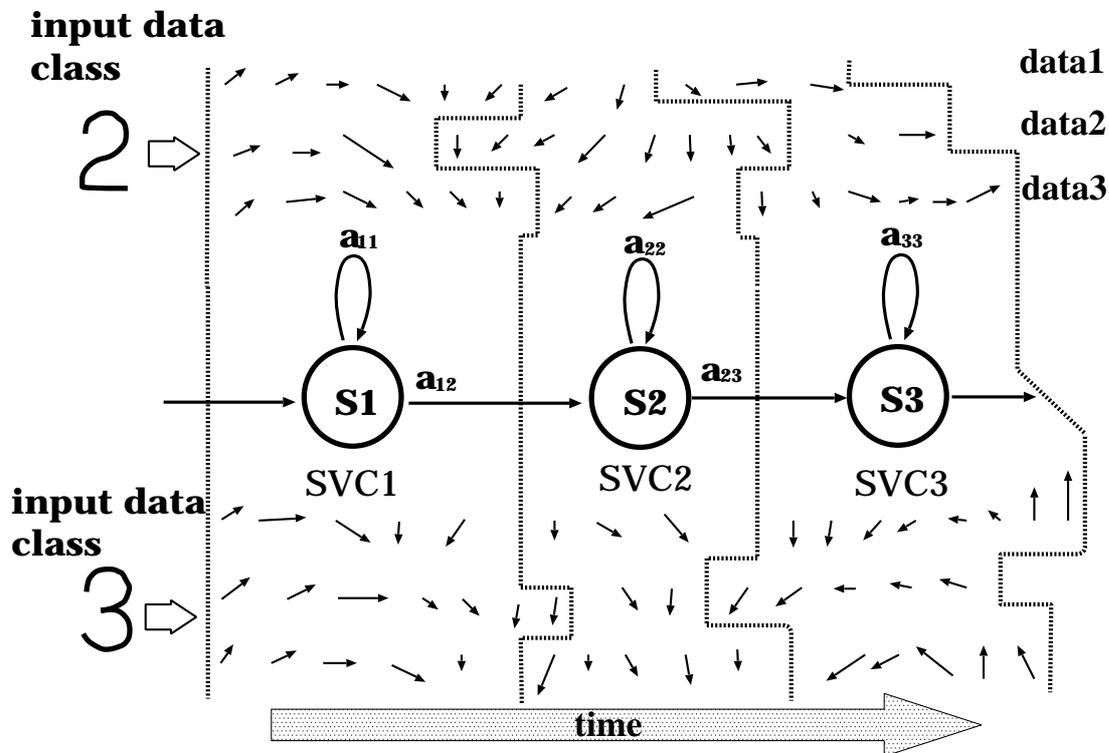


図 3.1: SVC 連鎖モデルによる手書き数字の認識

3.2.1 クラス帰属度

クラス帰属度は入力されたパターンが該当するクラスから生成される確かさを表す量として定義する。ここで、パターンが生成される確かさを最適分離超平面からの距離で表現できると仮定をし、クラス帰属度を分離超平面からの距離に比例するように定義する。つまり最適分離超平面から一方のクラスの側への距離が大きいと、そのクラスに対するクラス帰属度も大きいとする。

本研究では最適分離超平面からの距離尺度に識別関数の値を用いた、識別関数の値は図(3.2)のように最適分離超平面で0となり最適分離超平面からの距離に従い大きな値を持つようになる。しかし識別関数の値を直接扱ったのでは1つのSVC内では有効な距離尺度になるが、異なるSVCに共通した距離尺度ではない。そこで個々のSVCについてサポートプレーン間の距離を考慮してクラス帰属度を次式のように定義する。

$$F \left(g_i(\mathbf{x}) \cdot \frac{1}{\sqrt{\mathbf{w} \cdot \mathbf{w}}} \right) \quad (3.1)$$

ここで、 $g_i(\mathbf{x})$ は状態 i における SVC の識別関数 (式 (??)) であり、 $\frac{1}{\sqrt{\mathbf{w} \cdot \mathbf{w}}}$ はクラス間距離を表している。 $F(\cdot)$ は値域 $[0, 1]$ なる単調増加関数で、本研究では次式のようなシグモイド関数 (式 3.2) を用いた。

$$F(x) = \frac{1}{1 + e^{-Kx}} \tag{3.2}$$

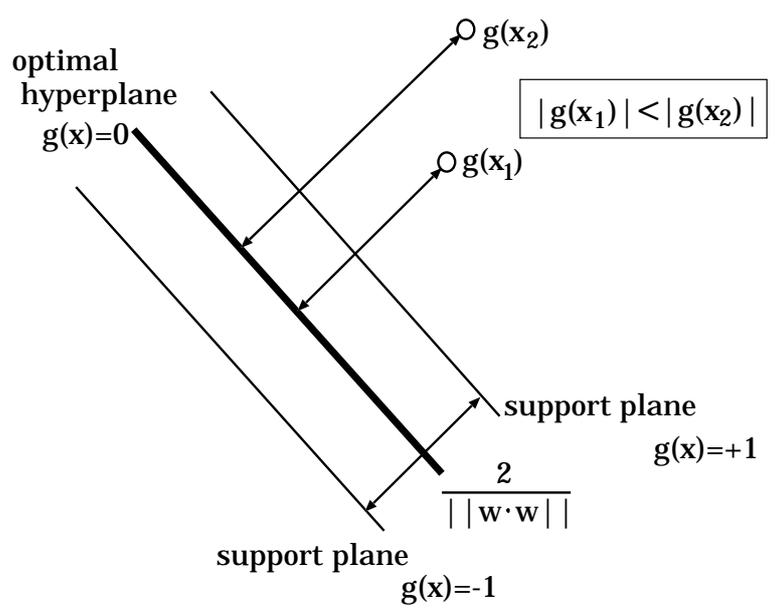


図 3.2: クラス帰属度

3.2.2 遷移の重み

遷移の重みは、SVC の遷移の容易さを表すパラメータである。CSVC では遷移の重み $a_{i,j}$ (i 番目の SVC から j 番目の SVC への遷移重み) を次式のように遷移の回数を数え上げることで求める。

$$a_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,k}} \quad k: \text{遷移可能な SVC の番号} \quad (3.3)$$

ここで $n_{i,j}$ は、学習データにおいて i 番目の SVC から j 番目の SVC への遷移が生じた回数である。

3.3 SVC 連鎖モデルの認識アルゴリズム

長さ n の入力時系列パターン $X = (x_1, \dots, x_n)$ が与えられたとき、パラメータセット Λ を有する状態数 S の SVC 連鎖モデルによる識別器の出力 $G(X, \Lambda)$ を次式で定義する。

$$G(X, \Lambda) = \max_{q_1, \dots, q_n} \prod_{t=1}^n a_{q_{t-1}q_t} F(g_{q_t}(x_t)) \quad (3.4)$$

ここで、 q_t は、時刻 t における状態番号 $1 \leq q_t \leq S$ を表わし、上式の右辺は可能な状態系列 (q_1, q_2, \dots, q_n) の中から、 $\prod_{t=1}^n a_{q_{t-1}q_t} F(g_{q_t}(x_t))$ が最大となる系列を探す処理を意味する。

また認識の段階においてクラス帰属度関数は SVC の識別基準と類似した次式についても提案する。

$$F(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3.5)$$

式 (3.5) は SVC が [目的クラス] と識別したフレーム数を数え上げることになる。上記の探索問題は HMM で用いられる Viterbi アルゴリズム [16] と同様の解法によって高速に最適解を求めることができる。Viterbi アルゴリズムによって最適状態系列を求めること、すなわち、入力時系列パターンを各状態に最適に割り振ることを Viterbi セグメンテーションとよぶ。

3.4 複数カテゴリの認識 (3 カテゴリ以上)

SVC は 2 クラスの識別器である、そのため CSVC も 2 カテゴリの識別器となる。しかし音声認識や文字認識などに代表される実際のパターン認識の問題は 3 カテゴリ以上の

多カテゴリ問題である。これまで SVC のような 2 カテゴリの識別器による多カテゴリ問題の対処法としては、[1 カテゴリ] 対 [多カテゴリ型]、対判別型、などがある、本研究では [目的の 1 カテゴリ] と [その他の全カテゴリ] を識別する CSVC を全てのカテゴリについて用意し最も高いスコアを出力したモデルを選択する (図 3.3)[1 カテゴリ] 対 [多カテゴリ] 型による識別を行う。

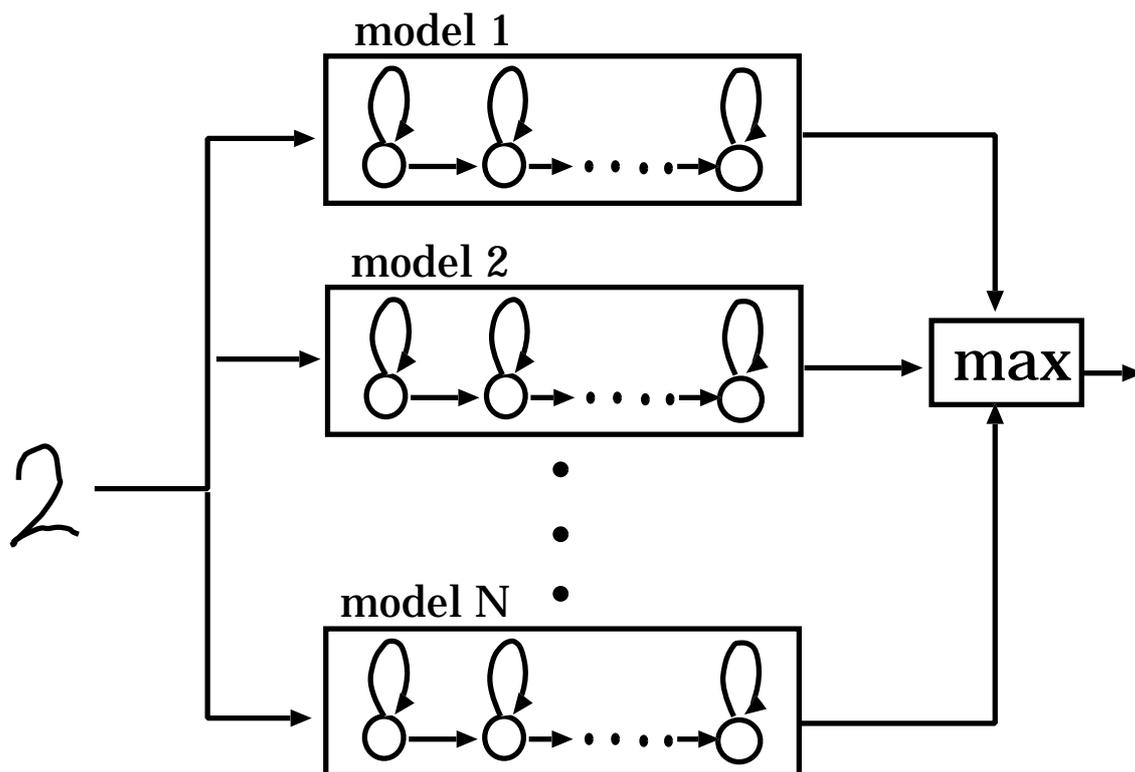


図 3.3: 複数カテゴリの認識

3.5 SVC 連鎖モデルの学習アルゴリズム

3.5.1 Viterbi アルゴリズム

CSVC は HMM などでも用いられている Viterbi アルゴリズムと同様の学習アルゴリズムを利用している、そこで Viterbi アルゴリズムの簡単な説明を行なう。

Viterbi アルゴリズムは、HMM の可能な状態遷移系列の中から最大の確率を与える遷移系列を探索するアルゴリズムであり、厳密に確率を求める方法と比較して、計算量が少ないにも関わらず認識性能は同等であることが確かめられている。

step 1 $\alpha(i, t) = \pi_q$: if q 初期状態

$\alpha(i, t) = 0$: otherwise

step 2 $t = 1, 2, \dots, T$ (入力フレーム) について step3,4 を実行

step 3 $q = 1, 2, \dots, Q_N$ (全ての状態) について step4 を実行

step 4 $\alpha(q, t)' = \max_i \{ \alpha(i, t)' \times a_{i,q} \cdot b_i(x_t) \}$

$L = \max_q \alpha(q, T)'$: if q 終了状態

ここで $a_{i,j}$ は遷移確率、 $b_i(x_t)$ は出力確率、 $\alpha(i, t)$ は時刻 t 、状態 i における確率である。入力された時系列データは系列長 T の $X = (x_1, x_2, \dots, x_T)$ である。

3.5.2 CSVC 学習アルゴリズム

状態数 S から成る CSVC の学習問題を、ここでは、次式に示すように各状態の SVC の最適化問題の評価関数 L_s , $s = 1, \dots, S$ の総和である L_a を最小化する問題として定式化する。

$$\text{Minimize } L_a = \sum_{s=1}^S L_s \quad (3.6)$$

$$\text{subject to } y_{si}(\mathbf{w}_s \cdot \mathbf{x}_{si} + b_s) \geq 1 - \xi_{si} \quad (3.7)$$

$$\mathbf{x}_{si} \in \{ \mathbf{x}_t | q(\mathbf{x}_t) = s \} \quad \text{for all } t, s, i$$

ここで、 $q(\mathbf{x}_t)$ は学習データ \mathbf{x}_t が属する状態の番号を、 \mathbf{x}_{si} および y_{si} は状態 s に属する i 番目の学習データの特徴量とクラス番号を意味する。評価関数 L_s は式 (2.4) と同様のものを用いる。

上記の最適化問題を解くには、学習データのセグメンテーション、すなわち、学習データと SVC との対応関係を予め決めておく必要がある。学習データの分割は、3.3節で示したように CSVC による認識処理 (Viterbi セグメンテーション) により決定できる、それには CSVC のパラメータが決定されている必要があるが、CSVC のパラメータは学習データの分割を決めなければならない。同様の問題に対する解法として HMM の学習法の一つである Viterbi アルゴリズムがあり、ここでは、次に示すように Viterbi 学習と類似した反復処理によるパラメータ推定を行う。

- step 1. カウンタを $\ell = 0$ とし、各学習資料を状態の数だけ等分割する (初期セグメンテーション)。
- step 2. 分割された学習データについて CSVC のパラメータセット Λ_ℓ を最適化問題 (3.6) に基づいて決定する。
- step 3. step 2 で決定した Λ_ℓ を用いて、学習資料の Viterbi セグメンテーションを行う。
- step 4. 予め定めた終了条件を満足していれば終了、そうでなければ $\ell \leftarrow \ell + 1$ として、step 2 へ。

HMM で用いられる Viterbi アルゴリズムと step 3 で行なわれるアルゴリズムが異なる点は、Viterbi アルゴリズムが確率モデルとして定式化されている HMM に適応されるのに対して、step 3 で行なわれるアルゴリズムはクラス帰属度が確率として定義されていないため確率モデルとしての定式化がなされていない CSVC に適応される点である。CSVC が確率モデルでないため学習アルゴリズムの収束性については十分な検討が必要である、また他のアルゴリズムによる学習法の検討も有効であると考えられる。

3.6 学習アルゴリズムの収束性

前節で示した学習アルゴリズムの収束性の証明はできていない。しかし以下の条件下であれば収束が可能である。

仮定 1 遷移の重み $\{a_{ij}\}$ が全て等しい (すなわち $a_{ij} = 0.5$)

仮定 2 各 SVC の学習誤りが無い

仮定 3 学習によりサポートプレーン間距離の大きな SVC からサポートプレーン間距離の小さな SVC へデータの遷移が生じる。

[証明] 状態 k に属する学習データ \mathbf{x}_t が Viterbi セグメンテーションにより $k + 1$ 番目の状態に移動した場合を考える。

まず、もし、 \mathbf{x}_t が状態 k においてサポートプレーン上の点 (すなわち、サポートベクトル) であれば、この点が隣りの状態 $k + 1$ に移動することによって、式 (2.4) で与えられる評価関数 L_k は減少する。サポートベクトルでなければ L_k は変化しない。一方、移動先の状態 $k + 1$ の SVC については以下のことが言える。

$$a_{k,k+1} F \left(\frac{g_{k+1}(\mathbf{x}_t)}{\sqrt{\mathbf{w}_{k+1} \cdot \mathbf{w}_{k+1}}} \right) > a_{k,k} F \left(\frac{g_k(\mathbf{x}_t)}{\sqrt{\mathbf{w}_k \cdot \mathbf{w}_k}} \right) \quad (3.8)$$

が成立している。仮定 1 から、

$$F \left(\frac{g_{k+1}(\mathbf{x}_t)}{\sqrt{\mathbf{w}_{k+1} \cdot \mathbf{w}_{k+1}}} \right) > F \left(\frac{g_k(\mathbf{x}_t)}{\sqrt{\mathbf{w}_k \cdot \mathbf{w}_k}} \right) \quad (3.9)$$

であり、式 (3.2)、(2.5)、および、仮定 2、3 から、

$$g_{k+1}(\mathbf{x}_t) > g_k(\mathbf{x}_t) \geq 1 \quad (3.10)$$

が成立する。したがって、 \mathbf{x}_t は状態 $k + 1$ の対応するサポートプレーンの外側に位置するため、状態 $k + 1$ の SVC 学習において 2 つのサポートプレーン間の距離 ($2/\sqrt{\mathbf{w} \cdot \mathbf{w}}$) を小さくする働きをしない。すなわち、 L_{k+1} は変化しない。したがって、CSVC 全体の評価関数 L_a は減少するか変化しないかのどちらかである。□

第 4 章

手書き数字文字の認識実験

提案した CSVC の識別性能を評価するために、手書き数字文字の認識実験を行った。

実験に用いた手書き数字文字データは、東京工業高等専門学校の手書き文字データベース TCT1 (筆者数 106 名)[19] 中の数字文字である。

実験に用いたフレーム特徴量は、文字の開始点を基準としたときの「相対座標ベクトル」と、その時間差分である「速度ベクトル」で、それぞれ 2 次元ベクトルである。速度ベクトルの時系列は手書き文字の認識に有効な特徴量であることが知られているが、各速度ベクトルの 2 次元ベクトル空間上の分布は、クラス間で大きくオーバーラップしている。そのため、非線形な識別境界を求める SVC の特徴を活かしにくい学習問題であると予想される。そこで、「速度ベクトル」のみの 2 次元ベクトルと、「相対座標ベクトル」も含めた 4 次元ベクトルの 2 種類のフレーム特徴量について実験を行う。また、今回の実験では SVC のカーネル関数には次式に示す次数 d の多項式を用いた。

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d \quad (4.1)$$

また、以下の 3 項目について異なる CSVC のモデルを作成し、比較実験を 4 次元特徴量の場合について行なった。

クラス帰属度関数

評価実験ではクラス帰属度関数として式 (3.5) を用いた場合と、式 (3.2) を用いた場合の比較実験は 4 次元特徴量の場合について行なっている (4.2.4 節)、本稿で報告する 2 次元特徴量による実験結果は式 (3.5) を用いたものである。

遷移の重み

遷移の重みは SVC 間の遷移の容易さを表すパラメータであり、同時に時系列データ

の時間方向の伸縮を表している。(3.6節)において示した学習アルゴリズムの収束を可能にする条件を満たすために最尤推定による遷移の重みと値を全て等しくした遷移の重みについて比較を行なった。比較実験の結果は4次元特徴量について(4.2.3節)に示した、2次元特徴量による実験は、全ての遷移の重みが等しいモデルを用いたものである。

誤り項の正規化

CSVCの学習においては、[1クラス]対[多クラス]の識別を行なうため認識対象クラスのデータ数(N_{+1})に対して「その他」クラスのデータ数(N_{-1})が遥かに多くなるため、式(2.4)の右辺第2項における誤分類の評価に偏りが生じる。そこで、次式のように各クラスに属するデータ数で正規化したものを用いた。式(2.4)と式(4.2)の比較実験は4次元特徴量の場合について行なっている(4.2.5節)。

$$L = \frac{\lambda}{2} \mathbf{w} \cdot \mathbf{w} + (1 - \lambda) \left[\frac{1}{2N_{+1}} \sum_{y_i=+1} \xi_i + \frac{1}{2N_{-1}} \sum_{y_j=-1} \xi_j \right] \quad (4.2)$$

4.1 2次元特徴による評価実験

SVCには経験的に決定されるパラメータとして、式(2.4)で導入した λ とカーネル関数のパラメータである d がある。これらのパラメータを決定するために予備実験を行なった。予備実験は学習用データとして各字種10文字(筆者10名:筆者1人のデータ量は1字種あたり1文字)の100文字のセットを3組使用し、評価用データとしては学習用データとは完全に異なる筆者10名による各字種10文字(筆者10名:筆者1人のデータ量は1字種あたり1文字)の100文字のセット5組みを用いた手書き数字認識実験である。

4.1.1 重み λ に関する検討

SVCは、経験的誤りの項とクラス間距離の逆数の項の和の最小化問題と定式化(式2.4)された。この定式化において2つの項の重みを表す定数 λ を導入した、予備実験では5回の反復学習を行なった結果、重み λ による最高認識率の変化を調査した。予備実験の結果をグラフ(4.1)に示した、この結果から2次元特徴量の場合 λ の値を'0.005'と決める。

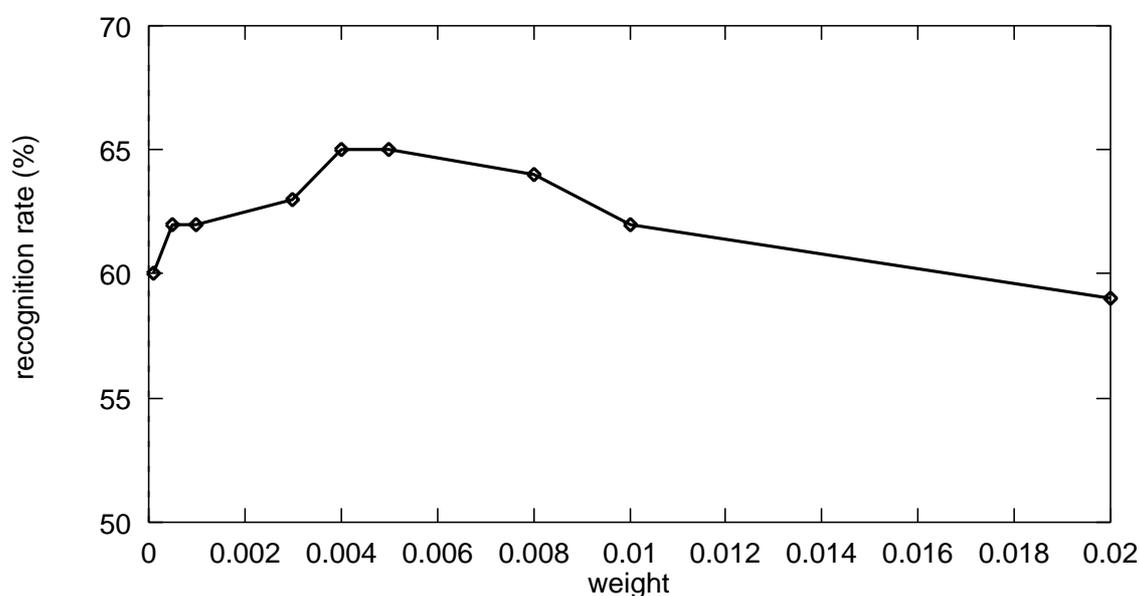


図 4.1: λ による認識率の変化(2次元特徴量)

4.1.2 カーネル関数の次数 d に関する検討

次数 d を決める予備実験も重み λ に関する予備実験と同様に、5 回の反復学習を行なった結果、次数 d による最高認識率の変化を調査した。予備実験の結果をグラフ (4.2) に示した、グラフの横軸は d であり縦軸は認識率である。この結果から 2 次元特徴量の場合 d の値を '5' と決める。

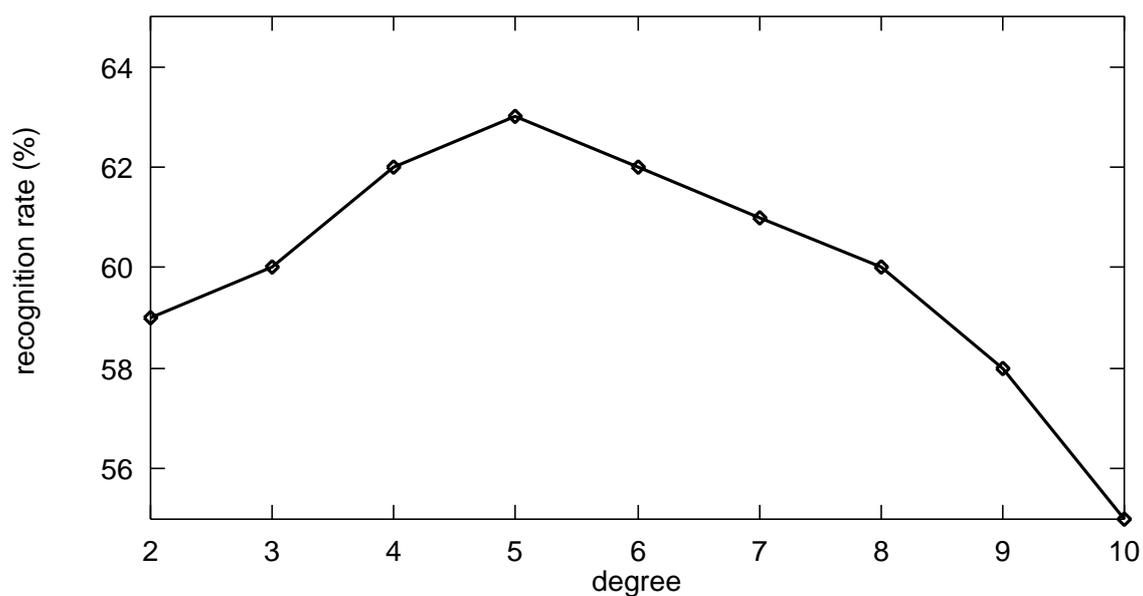


図 4.2: d による認識率の変化 (2 次元特徴量)

予備実験により求めた $\lambda = 0.005$ と $d = 5$ を使い、データ数を増やした実験を行なった。実験は学習データとして各字種 20 文字 (筆者 20 名 : 筆者 1 人のデータ量は 1 字種あたり 1 文字) の 200 文字のセットを 2 組使用し、評価用データとしては学習用データとは完全に異なる筆者 10 名による各字種 10 文字 (筆者 10 名 : 筆者 1 人のデータ量は 1 字種あたり 1 文字) の 100 文字のセット 3 組みによる手書き数字認識である。

4.1.3 反復学習の回数に関する検討

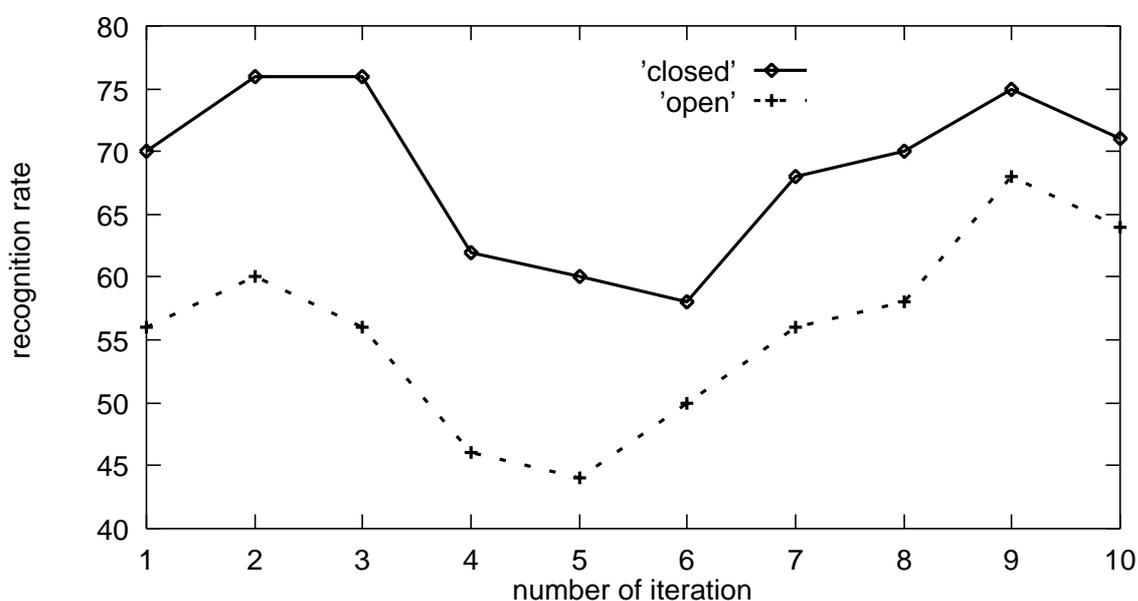


図 4.3: 学習曲線 (2 次元特徴の場合)

反復学習 (3.5 節) の回数と認識率との関係をグラフ 4.3 に示す。グラフの横軸は Viterbi 学習の反復回数、縦軸は認識率である。学習の反復により認識率の向上が若干見られるが、変動が大きい上に収束性は確認されなかった。原因としては、特徴量の次元数が 2 次元と低く、対象クラスとその他クラスが特徴空間上で大きくオーバーラップしており、超平面による分離が困難であるためだと考えられる。

4.1.4 HMM との性能比較

CSVC の性能比較のために HMM との認識率の比較を行なった (表 4.1)。使用した HMM は 4 状態、単一正規分布型のモデルで、10 回の Viterbi 学習を行った。CSVC の認識率は反復学習において最も高い認識率を示している。

この比較実験では CSVC の認識性能は HMM と比べて極めて低いことが示された。CSVC の認識性能が HMM に劣る理由としてはデータの次元数が低く特徴空間上でデータのオーバーラップが大きいことが考えられる。

HMM に対して SVC のような識別学習を行なう識別器では対立クラスのデータの分布形状にも影響を受けるため、クラス間のオーバーラップが大きい場合クラス間の分離境界の決定が困難になることが予測される。このような問題に対して SVC はデータを高次元空間に写像して分離境界を決定する特徴を持っているため本来分離が困難であるデータに対しても正しい分離境界を決定することができる、しかし今回の問題ではデータの次元数が 2 次元と低いために高次元空間へ写像する特徴が十分に活かされなかったのではないかと推測される。更に反復学習で用いたアルゴリズムとクラス帰属度の整合性にも問題があったと考えられる、この問題については (5) で検討を行なう。

表 4.1: (2 次元特徴量の場合)

	training(%)	test(%)
HMM	97	91
CSVC	75	68

4.2 4次元特徴による評価実験

2次元特徴量の場合と同様に、パラメータ λ と d に関する予備実験を行なった。予備実験は2次元特徴量の場合と同様に学習用データとして各字種10文字(筆者10名:筆者1人のデータ量は1字種あたり1文字)の100文字のセットを3組使用し、評価用データとしては学習用データとは完全に異なる筆者10名による各字種10文字(筆者10名:筆者1人のデータ量は1字種あたり1文字)の100文字のセット5組みによる手書き数字認識実験である。

4.2.1 重み λ に関する検討

パラメータ λ に関する予備実験では5回の反復学習を行なった結果、重み λ による最高認識率の変化を調査した。この実験の結果をグラフ(4.4)に示す、グラフの横軸は重み λ であり縦軸は認識率である。この結果から λ の値を'0.01'と決定した。

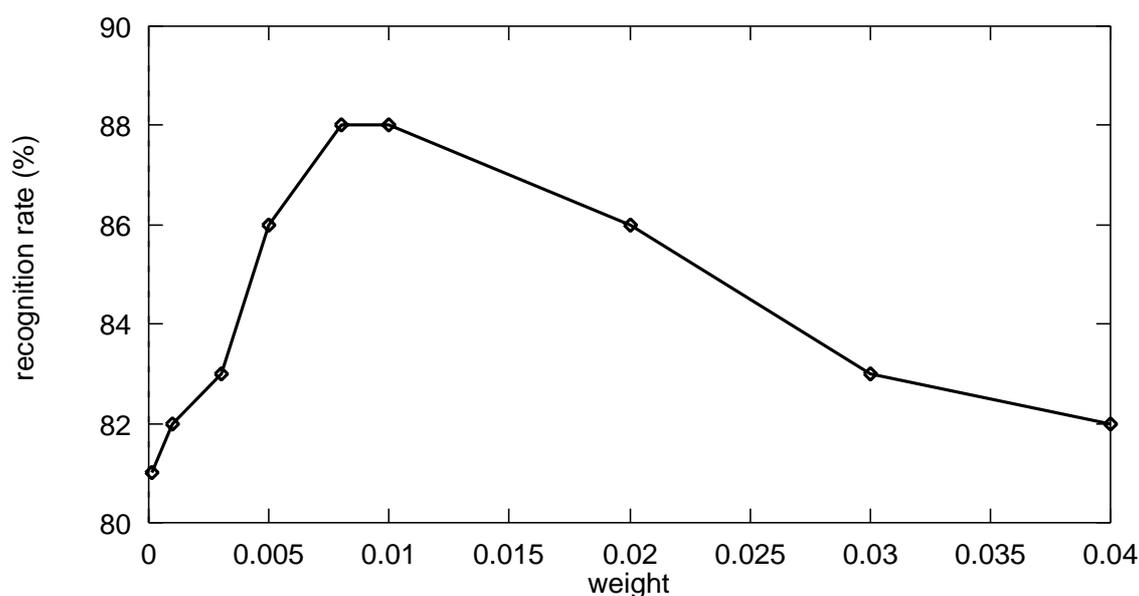


図 4.4: λ による認識率の変化(4次元特徴量)

4.2.2 カーネル関数のパラメータ d に関する検討

次数 d を決める予備実験も重み λ に関する予備実験と同様に、5回の反復学習を行なった結果、次数 d による最高認識率の変化を調査した。予備実験の結果をグラフ(4.5)に示

した、グラフの横軸はパラメータ d であり縦軸は認識率である。この結果から 2 次元特徴量の場合 d の値を '13' と決める。

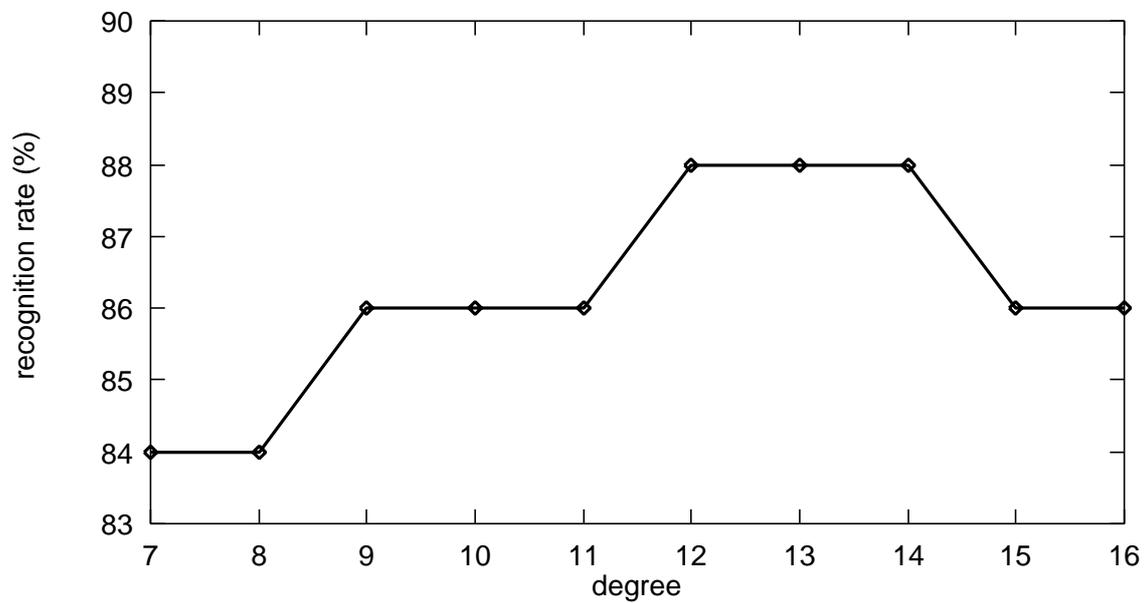


図 4.5: d による認識率の変化 (4 次元特徴量)

4次元特徴量については誤り項の正規化と遷移重みに対する比較実験を行った、比較実験は学習データとして各字種10文字(筆者10名:筆者1人のデータ量は1字種あたり1文字)の100文字のセットを3組使用し、評価用データとしては学習用データとは完全に異なる筆者10名による各字種10文字(筆者10名:筆者1人のデータ量は1字種あたり1文字)の100文字のセット5組みを使用した、手書き数字認識である。パラメータは予備実験により求めた $\lambda = 0.01$ と $d = 13$ を用いる。

4.2.3 遷移の重みに関する検討

前章で述べた学習アルゴリズムが収束可能な条件の1つに[遷移の重みが全て等しい($a_{i,j} = 0.5 \quad \forall i, j$)]があった、そこでこの条件に従い遷移の重みを全て等しい(0.5)場合と最尤推定により求めた場合についての比較を行なった図(4.6)。横軸は反復学習の回数であり縦軸は認識率である。

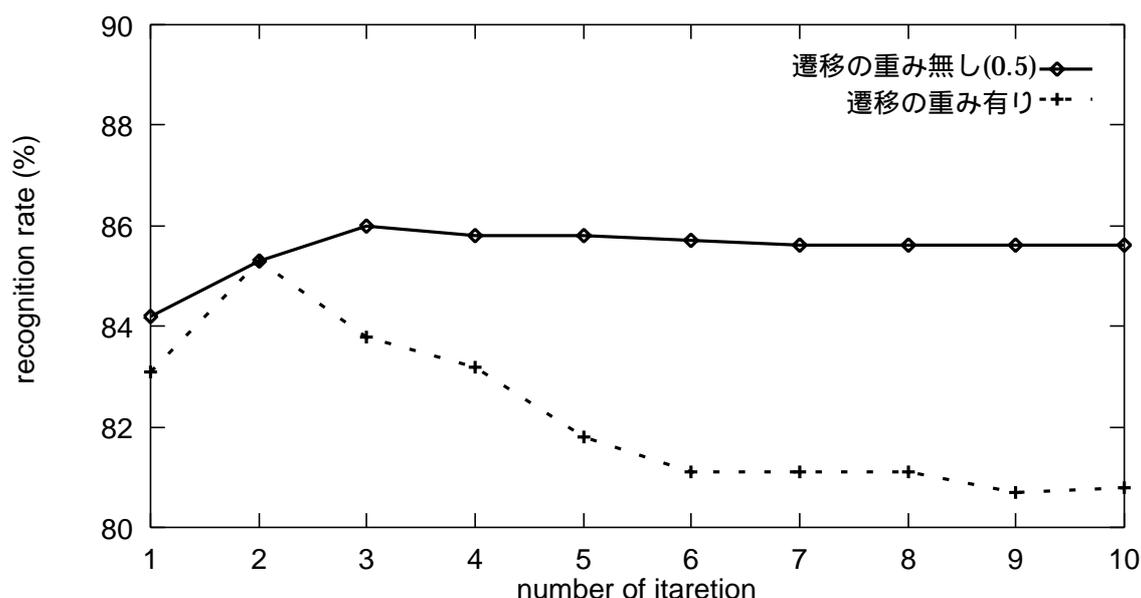


図 4.6: 遷移重みによる認識率の変化

図(4.6)より遷移の重みが等しい場合の最も高い認識率は86%であり最尤推定の場合は85.3%であった。最も高い認識率で比較すると2つのモデルの差は小さいが反復学習に対して安定して高い認識率を示した点も考慮して、以降の実験では遷移の重みの値は全て0.5となるモデルで行なう。

4.2.4 クラス帰属度関数に関する検討

先に示したように本研究ではクラス帰属度関数として、データが最適分離超平面のどちら側にあるかで $[0 \text{ or } 1]$ の判定をするステップ関数(式(3.5))と、データと最適分離超平面の距離を考慮するシグモイド関数(式(3.2))の2つを提案した。2つのクラス帰属度関数について比較実験を行ない、結果をグラフ(4.7)に示した、横軸は反復学習の回数であり縦軸は認識率である。

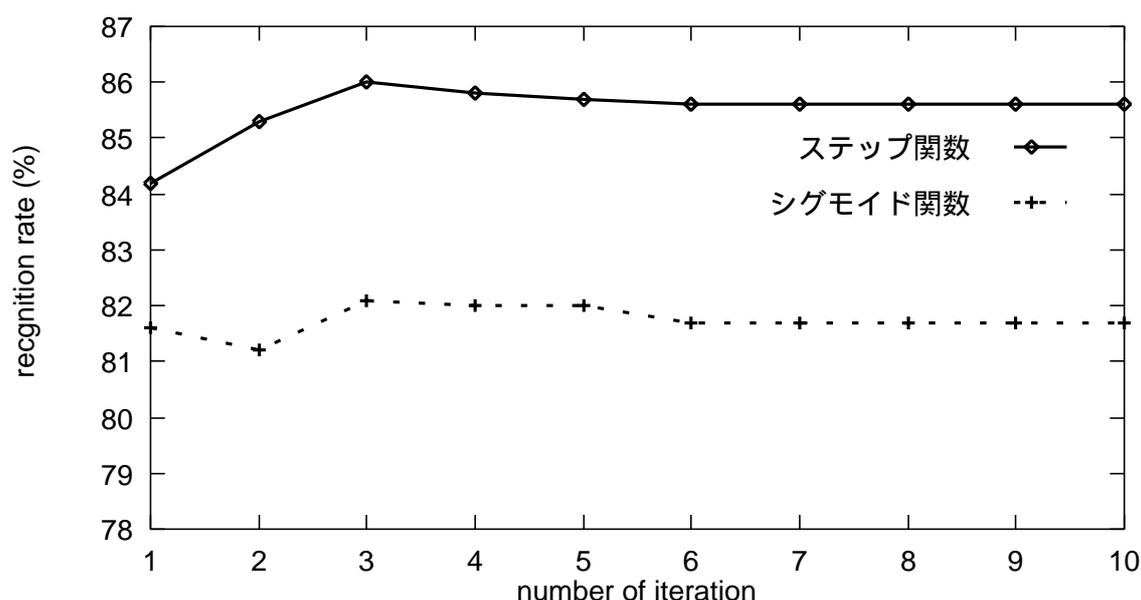


図 4.7: クラス帰属度による認識率の変化

グラフ(4.7)よりステップ関数を用いたモデルの最も高い認識率は86%でありシグモイド関数を用いたモデルの最も高い認識率は82.1%である。ステップ関数のモデルが高い認識性能を示した理由としては、SVCが識別境界のどちらの側かにデータが存在するか(識別関数の符号)だけで識別を行なうため識別境界からの距離を直接識別に利用する手法はあまり有効でないと考えられる。

4.2.5 誤り項の正規化法に関する検討

式 (4.2) で定義した学習誤りに対するデータ数の正規化を行なった場合と、行なわない場合 (式 (2.4)) との認識率の比較をグラフ (4.8) に示す、横軸は反復学習の回数であり縦軸は認識率である。グラフから、正規化を行なった場合の最高認識率が 85.3% であるのに対して、正規化を行わない場合の最高認識率は 81.2% であり、各クラスに含まれるデータ数で正規化を行った評価式 (4.2) を用いた方が良いことが分かる。

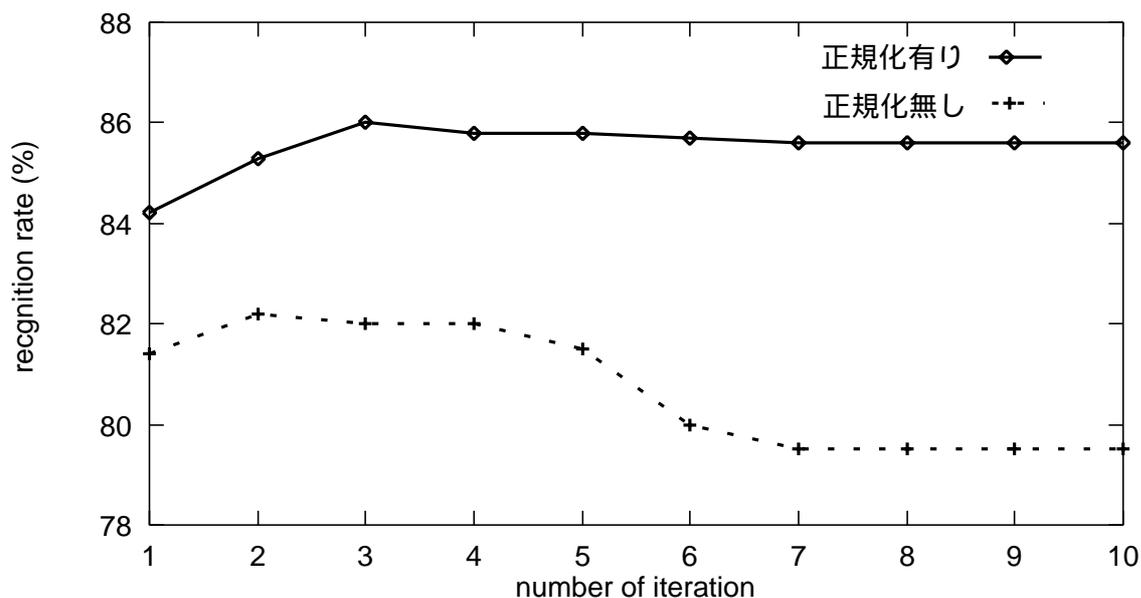


図 4.8: クラスデータ数による正規化の効果

4.2.6 HMM との性能比較

予備実験の結果から SVC のパラメータを $\lambda = 0.01$, $d = 13$ とし、遷移の重みを全て 0.5、クラス帰属度関数にステップ関数、誤り項の正規化を行なった場合の CSVC を用いた、HMM は特徴量 2 次元の場合と同様、4 状態単一正規分布型のモデルである。また CSVC は 3 回程度の反復学習により認識率が収束するため、反復学習を 3 回行なった CSVC と、Viterbi 学習を 10 回行なった HMM による比較実験を行なった。認識率を表 4.2 に示す。

実験はこれまでと同じ手書き数字データで行ない、フレーム特徴量は 4 次元である。実験は学習データとして各字種 40 文字 (筆者 40 名 : 筆者 1 人のデータ量は 1 字種あたり 1 文字) の 400 文字のセットを 2 組使用し、評価用データとしては学習用データとは完全に異なる筆者 10 名による各字種 10 文字 (筆者 10 名 : 筆者 1 人のデータ量は 1 字種あたり 1 文字) の 100 文字のセット 5 組みによる手書き数字認識である。

表 4.2 に CSVC、HMM の最も高い認識率を記載した、特徴量の次元数を 2 次元から 4 次元に増やすことにより、CSVC は HMM と同程度の認識精度が得られている。ただし、学習データが少ないため両者の厳密な比較を行なうには更に大規模なデータベースを用いた実験が必要があるので、現時点では優劣に関する結論は出せない。

表 4.2: HMM との比較 (4 次元特徴の場合)

	training (%)	test (%)
HMM	94.3	92.6
CSVC	95	91.6

第 5 章

実験の考察

本章では 4 章で行なった実験結果について、学習の収束性などを中心に考察を行なう。

5.1 認識率に関する考察

手書き数字認識による評価実験では、フレーム特徴量 2 次元の場合 CSVC の認識性能は HMM の認識性能を大幅に下回るものであった。原因としては特徴量空間上でクラス間のオーバーラップが大きかったにも関わらずデータの次元数が低かったためだと考えられる。

これに対してフレーム特徴量が 4 次元の場合の評価実験では認識率に大幅な向上が見られた、これは次元数が増えたことで特徴空間上でクラス間のオーバーラップが小さくなり更に高次元空間への非線形写像の効果が現れたためだと考えられる。さらに特徴量 4 次元の場合は反復学習により認識率の収束性が見られる、現在のところ学習アルゴリズムの収束性はある特定の条件下でのみ証明できるが今回の実験で満たされた条件は遷移の重みに関してだけであり、学習誤りとクラス間距離に関する条件は満たされていないと考えられる。そのため CSVC の認識率が一定の値に収束してはいるが、学習により収束した状態が定式化の解となっているのか詳細な調査が必要である。

次節では CSVC の収束性についての考察を行なう。

5.2 CSVC の収束性に関する考察

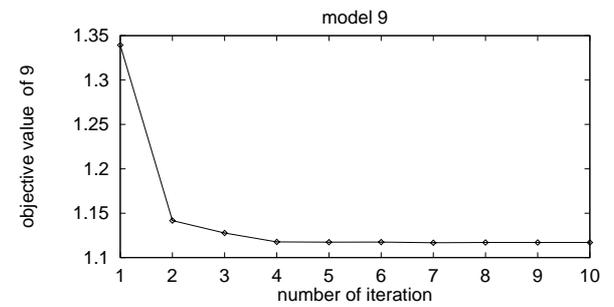
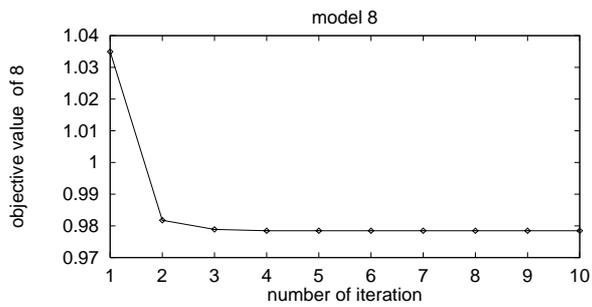
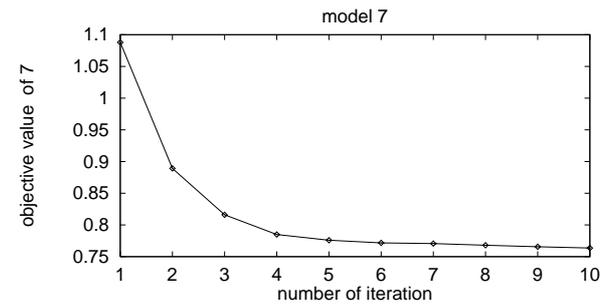
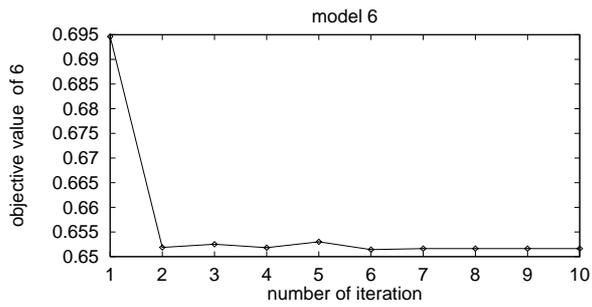
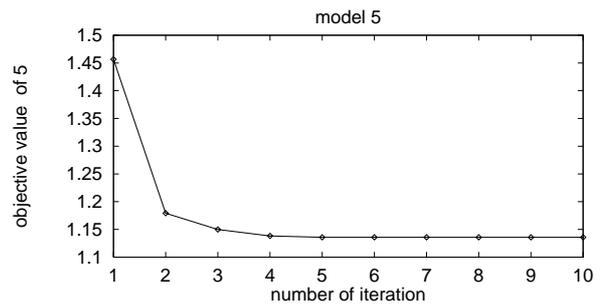
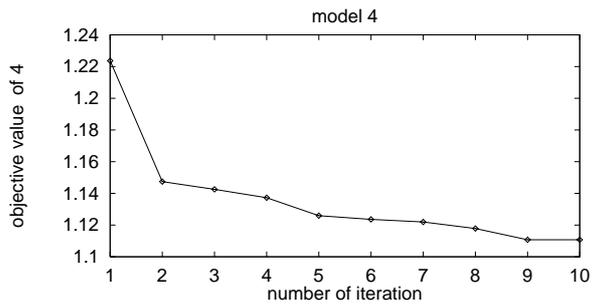
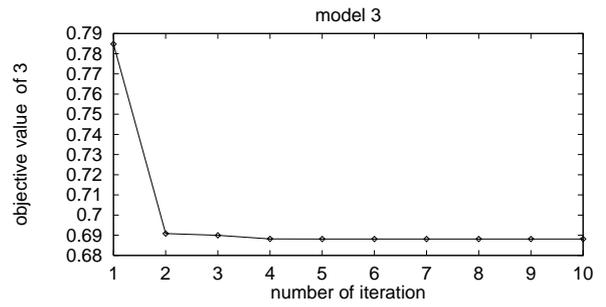
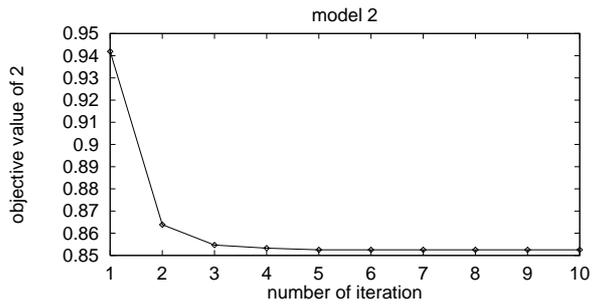
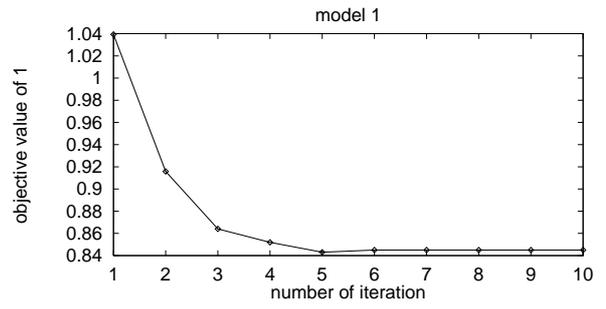
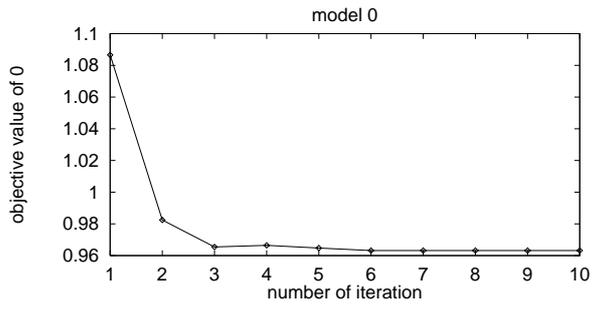
CSVC の収束性を学習の定式化で用いた評価関数値から調査した、次項に数字 '0 ~ 9' に対応した CSVC のモデル 'model 0 ~ model 9' の反復学習による評価関数値の変化を示す

グラフを掲載した。CSVC の評価関数は式 (3.4) で定義したが掲載したグラフの縦軸は評価関数の双対問題に相当する式 (2.11) であるため学習の定式化における最小化問題はこの場合最大化問題となり、横軸の反復学習が進行するに従い縦軸方向に増化することが望ましい。

しかし全てのモデルにおいて次項のグラフから評価関数は単調減少であることが示されている。

CSVC は学習データを分割し、分割されたデータごとに個々の SVC が対立するカテゴリのデータとの間に識別境界を構築する。CSVC の反復学習は個々の SVC による識別が最も容易になる最適のような学習データの分割を決定することを目的としている。しかし実験により CSVC の評価関数は学習の反復により値が悪くなることが確認された、これは反復学習により分割された学習データは個々の SVC にとって反復学習前よりも識別が困難な学習データとなるためである。実験結果の考察から、CSVC が反復学習により収束した状態は学習の目的とする状態でないと考えられる。

次節で本研究で行なった学習アルゴリズムの問題点について述べる。



5.3 学習アルゴリズムに関する検討

学習データの分割は Viterbi アルゴリズムと同様のアルゴリズムによる分割を行なう、HMM における Viterbi アルゴリズムとの相違点は CSV C の場合 HMM の出力確率の代わりにクラス帰属度を用いる点である。ここでは実験に用いたモデルと同様の、遷移の重みが全て等しモデルについて学習アルゴリズムの検証を行なう。

5.3.1 学習アルゴリズムの問題点

実験の考察により反復学習により、定式化における最大化問題の評価関数が減少していることが分かった、原因としては学習データの分割 (Viterbi 分割) の基準と SVC の学習基準の整合性がとれていなかったためだと考えられる。

学習データの分割 (Viterbi 分割) 基準は、データとの距離が最大になる最適分離超平面を構成する SVC の学習データとなるように分割を行なうことである。これに対して SVC の学習基準は、式 (2.4) で示したように学習誤りを減らし同時にクラス間距離を大きくすることであり、式 (2.4) の最小化を行なうことである。そして、この学習データの分割基準から与えられた学習データにより、SVC の評価関数である式 (2.4) が減少するとは限らない、これが学習データ分割と SVC の 2 つの学習基準の合いだの不整合性である。

(3.6) 節では、学習アルゴリズムの収束性についての証明を行なった。ここでは 3 つの条件が成立した場合についての証明であり、この 3 条件が成立する場合にのみ学習データ分割と SVC の 2 つの学習基準間の整合性が満たされることになる。収束性の証明を行なう際に必要であった 3 つの条件は、言い替えると、データの分割により SVC 間でデータが移動した場合に、移動した先の SVC にとってそのデータが誤り無く学習でき更にサポートベクターに選ばれないための条件となる。実験により全てのモデルで目的とした収束性が見られなかったのは収束の条件を満たすことなく、分割により移動したデータは移動先の SVC にとって学習誤りのデータであったり、サポートベクターに選ばれるものが少なくなかったためだと考えられる。

収束不可能な具体例

ここでは簡単な例として線形分離可能なデータを線形 SVC が 2 個連鎖した CSV C により識別する問題において、収束の条件が満たされないため、分割により移動したデータが SVC の再学習でサポートベクターに選ばれる場合について CSV C の評価関数 (式 (3.6)) の変動を考察する。

図 (5.1) では反復学習前の特徴量空間上のデータ分布と 2 つの SVC により構成された

超平面を表している、薄いグレーの分布 X は対立カテゴリーのデータの分布であり、' 分布 1' とある楕円が 1 番目の SVC の学習データ (目的クラス)、同様に ' 分布 2' とある楕円が 2 番目の SVC の学習データ (目的クラス) である。

この状態から Viterbi アルゴリズムによる学習を開始する。図 (5.1) のような分布であれば分布 1、分布 2 をそれぞれを 1 つの SVC に分割しているため、HMM と同様に CSVC も時系列データを定常信号源の連鎖と考えているので、現時点で最適な学習データの分割が行なわれていると考えられる、したがって学習によりこの分割は変化すべきでない。しかし最適分離超平面からの距離を基準にデータの分割を行うと最適のデータ分割が変化する可能性がある、ここで SVC1、SVC2 により構成された最適分離超平面と特徴空間上の任意の点の距離をそれぞれ h_1, h_2 とする。図 (5.1) の分布 1 の任意のデータについて 2 つの最適分離超平面との距離を比較すると $h_1 < h_2$ となる分布領域が存在する。クラス帰属度はデータと最適分離超平面との距離に対応しているため、 $h_1 < h_2$ となる領域に存在するデータは SVC1 に対するクラス帰属度よりも SVC2 に対するクラス帰属度の方が大きくなる、反復学習のアルゴリズムはクラス帰属度のより大きな SVC にデータを分割するため、 $h_1 < h_2$ なる領域のデータは SVC2 の学習データに移動する。

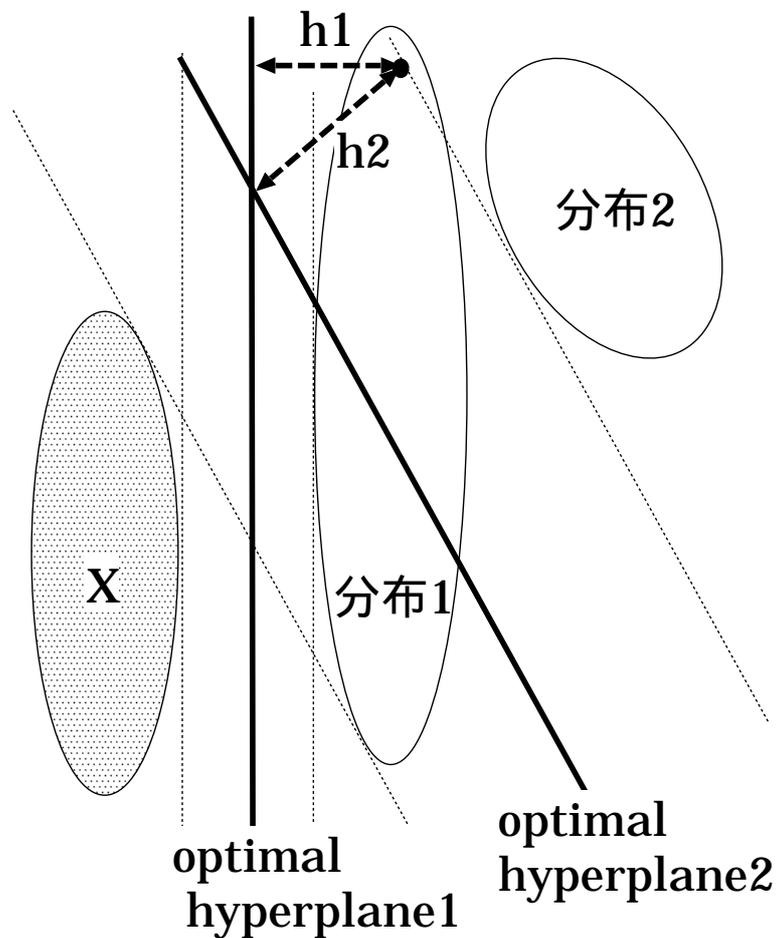


図 5.1: 学習前

データが移動した後に SVC の再学習を行う、その結果を図 (5.2) に示す。図 (5.2) では SVC2 が構成した最適分離超平面とサポートプレーンだけを表示した、SVC1 の最適分離超平面は図 (5.1) にあるものと変化しないのでここでは省略してある。図 (5.2) で濃いグレーの部分が入った部分が SVC2 の学習データの分布である、学習により SVC2 の学習データは分布 2 と分布 1 の一部を加えたものになっている。SVC1 から SVC2 に移動してきた分布 1 のデータは、再学習前の SVC2 にとっては誤り ($g(x) < 1$) となるデータであるので収束条件を満たしていない、そのため移動したデータは SVC2 の再学習によりサポートベクターに選ばれている。この時 SVC2 により構成される 2 つのサポートプレーン間の距離 (M') は学習前の距離 (M) よりも減少している ($M > M'$)。

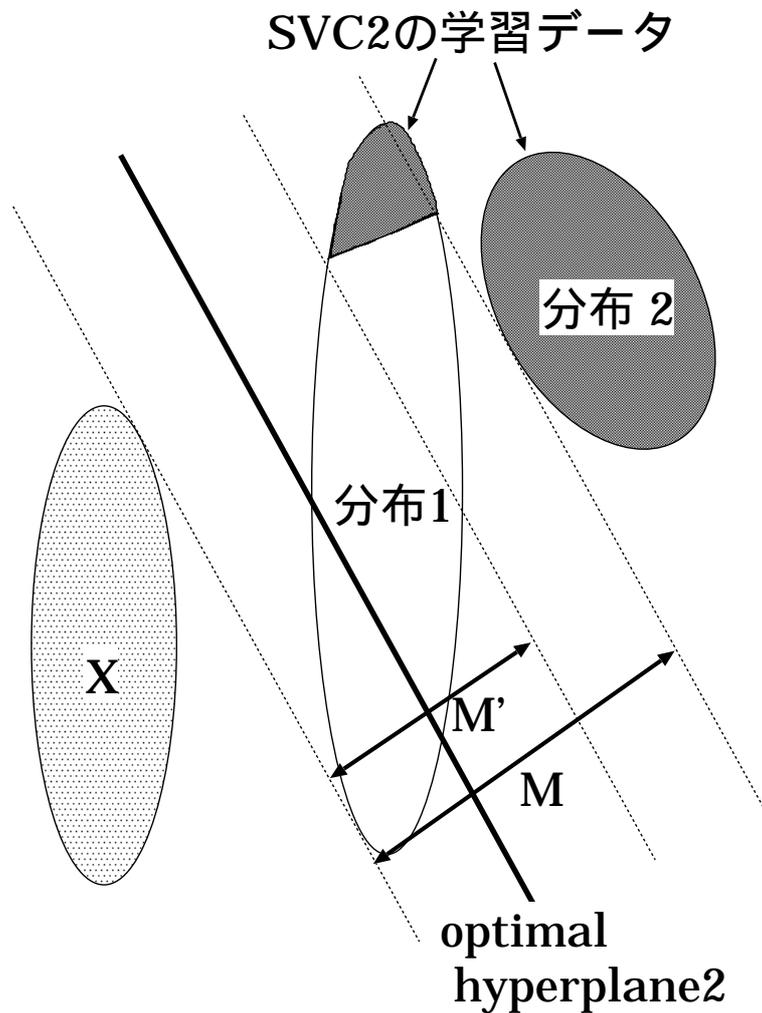


図 5.2: 学習後

ここで取り上げた例は、学習誤りが無いので SVC の評価関数はサポートプレーン間の距離にのみ依存しているため、CSVC の評価関数も同様にサポートプレーン間の距離にのみ依存する式 (5.1)。

$$\text{Minimize} : L = \frac{1}{2} \mathbf{w}_1 \cdot \mathbf{w}_1 + \frac{1}{2} \mathbf{w}_2 \cdot \mathbf{w}_2 \quad (5.1)$$

$$\text{subject to} : y_{si}(\mathbf{w}_s \cdot \mathbf{x}_{si} + b_s) = 1 \quad s : 1 \text{ or } 2 \quad (5.2)$$

ここで式 (5.1) の右辺第 1 項は SVC1 のサポートプレーン間距離の逆数であり第 2 項はサポートプレーン間距離の逆数である。学習による SVC1 のサポートプレーンの変化は無

いが、SVC2のサポートプレーンは移動しサポートプレーン間の距離が減少しているため次式のような関係にある。

$$\underbrace{\frac{1}{2}w_2 \cdot w_2}_{\text{学習前}} \leq \underbrace{\frac{1}{2}w'_2 \cdot w'_2}_{\text{学習後}} \quad (5.3)$$

この結果 CSVC の評価関数式 (5.1) は学習により増化するのでこのアルゴリズムでは CSVC の定式化における最適化問題の解が求まらないことになる。

ここでは反復学習により評価関数が増化する問題に対して 1 つ簡単な例を用いて説明を試みた、この問題の原因についてはここで示した例だけでなく他にも存在する可能性はある。しかし学習の定式化の最適化問題を解くためには少なくともここで明らかになった原因を解決する必要がある。

5.4 学習アルゴリズムの改善

本研究が提案した CSVC は時系列データを分割し、分割されたデータについてそれぞれ異なる SVC が対立カテゴリーのデータとの識別境界を構成している。CSVC は Viterbi アルゴリズムと同様のアルゴリズムにより CSVC の評価関数 (式 (3.4)) を最小にする最適な学習データの分割を求めようとしたが、提案した学習アルゴリズムでは最適な学習データの分割を求めることが困難であることが分かった。遷移確率と出力確率から Viterbi アルゴリズムによってパターンの生成される確率である尤度を求めている HMM と本研究で提案した CSVC の相違点は、HMM が各状態に出力確率を与えているのに対して CSVC が各 SVC にクラス帰属度を与えている点である。しかしクラス帰属度は HMM における出力確率のように、該当するクラスからパターンが生成される確かさを表すような量ではないと考えられる。クラス帰属度は最適分離超平面からの距離から定義したが、CSVC の学習アルゴリズムにより、入力データに対してより大きなクラス帰属度出力する SVC へ移動したとしても、再学習により移動先の SVC の評価関数が改善される保証はなく、特定の条件下でのみ評価関数の改善が保証される。問題解決のためには 2 つの方法が考えられる、1 番目の方法としてはクラス帰属度の定義の変更、2 番目の方法としてはクラス帰属度の定義は変更せずに学習データの分割を求めるアルゴリズムの変更である。

また SVC の評価関数である式 (2.4) を直接 Viterbi アルゴリズムと同様のアルゴリズムの評価量とすることも考えられるが、これはデータの移動を考える度に SVC の再学習が必要となり計算時間の面から非現実的であり実装は困難であると考えられる。

本研究で提案したアルゴリズムは収束性を十分に証明することができなかったが、次節で CSV C の評価量を変更して考察を行なった。

5.5 モデルの評価法に関する考察

これまでの議論では CSV C を式 (3.6) により評価していた。しかし、この評価関数と学習データの分割を決定するアルゴリズムとの関係が明確でない、そこでスコア (式 (3.4)) を用いた CSV C の評価を行なった。学習データの分割はスコアの最大化を基準に行なわれているので、学習の定式化を式 (3.6) の最小化問題とするよりもスコアの最大化問題と定義する事がより整合性があると考えられる。ただし、学習の定式化をスコアの最大化問題としても学習アルゴリズムも収束性については証明されていない。

ここでは学習・認識のアルゴリズムはこれまでと同様に第 3 章で示したアルゴリズムを用い、遷移の重みが全て等しく、クラス帰属が異なる 3 つのタイプを用いた実験を行ない、それぞれにつき反復学習によるスコアの変化を調査した。小量データの実験であるがこの実験結果から収束性の確認を試みる。

実験に用いたデータは 0~9 までの 10 種の手書き数字であり、特徴量は「相対位置座標」と「速度ベクトル」を合わせた 4 次元である。学習データには各字種 10 文字 (筆者 10 名 : 筆者 1 人当たりのデータ量は 1 字種 1 文字) の 100 文字を用い、評価用データとしては学習用データとは完全に異なる筆者 10 名による各字種 10 文字 (筆者 10 名 : 筆者 1 人当たりのデータ量は 1 字種 1 文字) の 100 文字をセット 5 組を用いた。

5.5.1 実験 1

実験 1 では、クラス帰属度として SVC の識別関数をシグモイド関数に代入した次式を用いて行なった。

$$F(g_i(\mathbf{x})) = \frac{1}{1 + \exp(g_i(\mathbf{x}))} \quad (5.4)$$

このクラス帰属度による認識率は学習用データに対して 99%であり、評価用データに対しては 84.5%であった。次のグラフに反復学習による認識率の変化を示す。

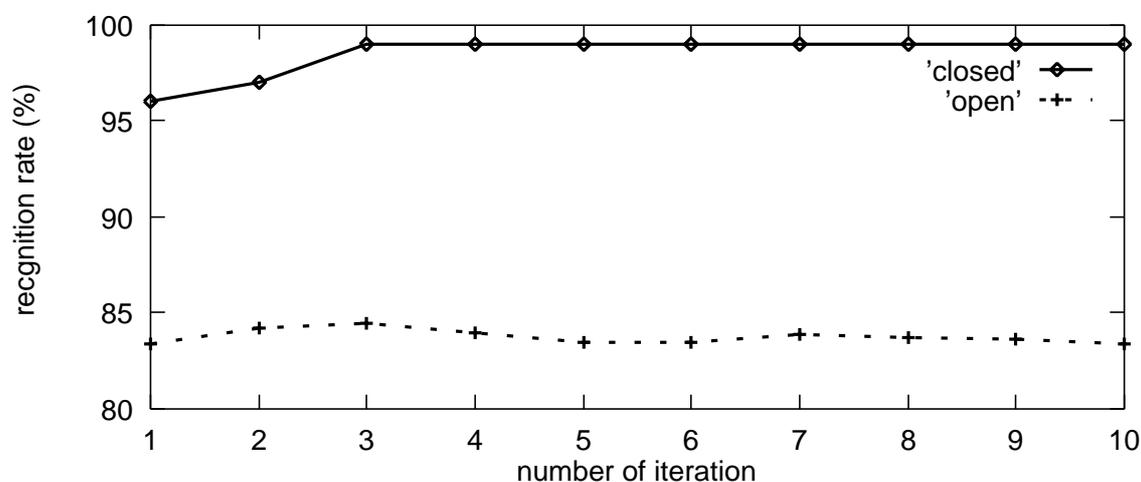
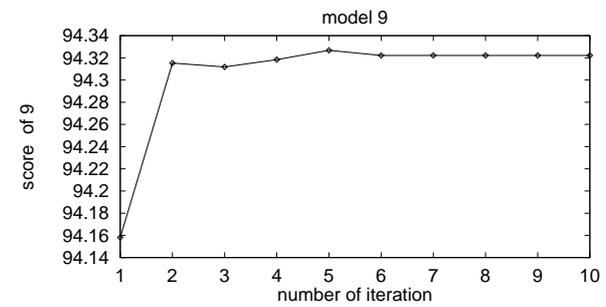
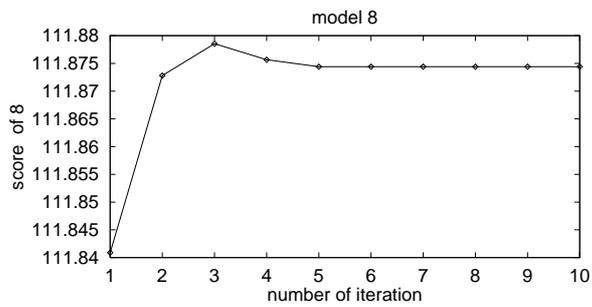
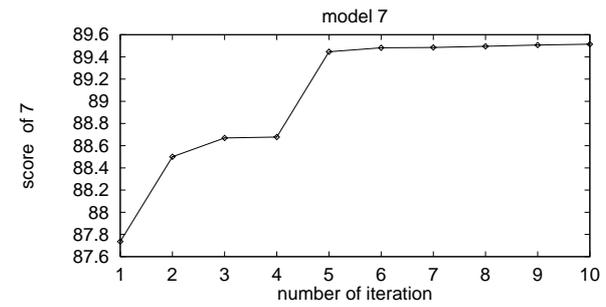
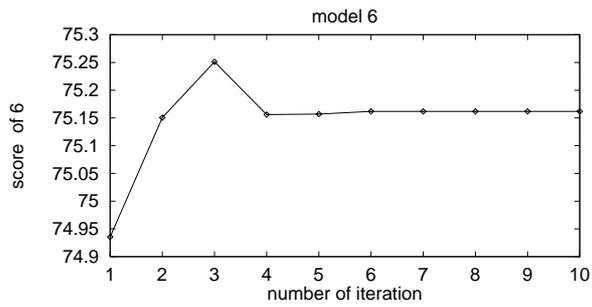
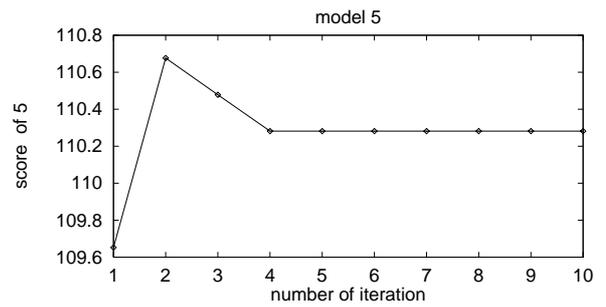
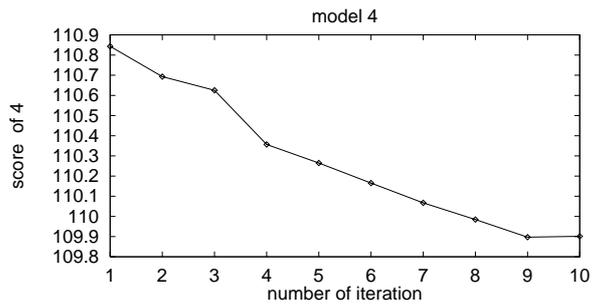
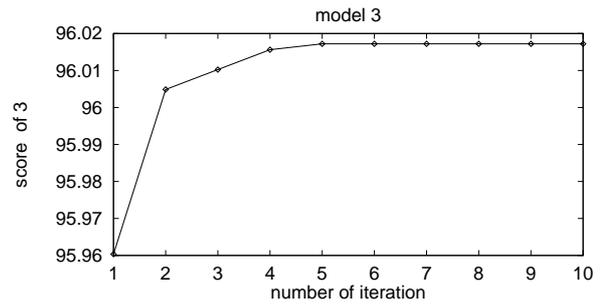
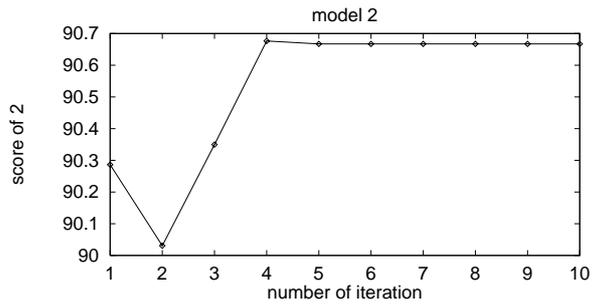
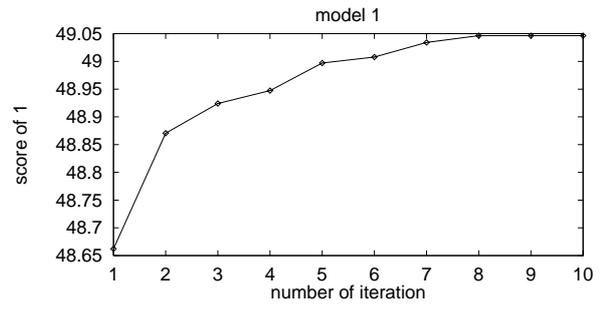
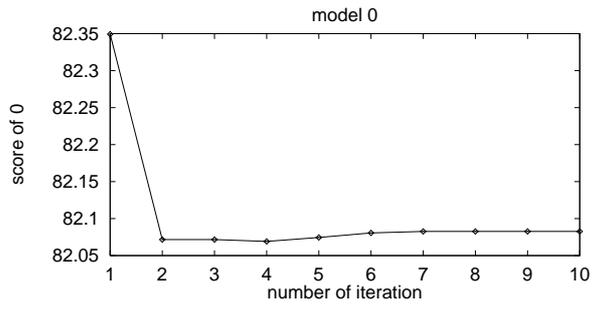


図 5.3: 識別関数によるクラス帰属度の認識率

次項に 0~9 までの各モデル (model0 ~ model9) の反復学習によるスコアの変化を表したグラフを掲載した、グラフの横軸は反復学習の回数であり縦軸はスコアである、スコアは各モデルに認識対象になっているカテゴリの学習データを入力したときの、出力の平均である。

これらのグラフから、model0 と model4 以外ではスコアが学習により増化する傾向が見られるため、スコアから CSVC の評価を行なうと実験的に収束性が確認できる。



5.5.2 実験 2

実験 2 では、クラス帰属度として第 4 章と同様の SVC の分離境界からの距離を利用した次式を用いた。

$$F(g_i(\mathbf{x})) = \frac{1}{1 + \exp(d_i(\mathbf{x}))} \quad (5.5)$$

$$d_i(\mathbf{x}) = g_i(\mathbf{x}) \cdot \frac{1}{\sqrt{\mathbf{w} \cdot \mathbf{w}}} \quad (5.6)$$

このクラス帰属度による認識率は学習用データに対して 99%であり、評価用データに対しては 85%であった。次のグラフに反復学習による認識率の変化を示す。

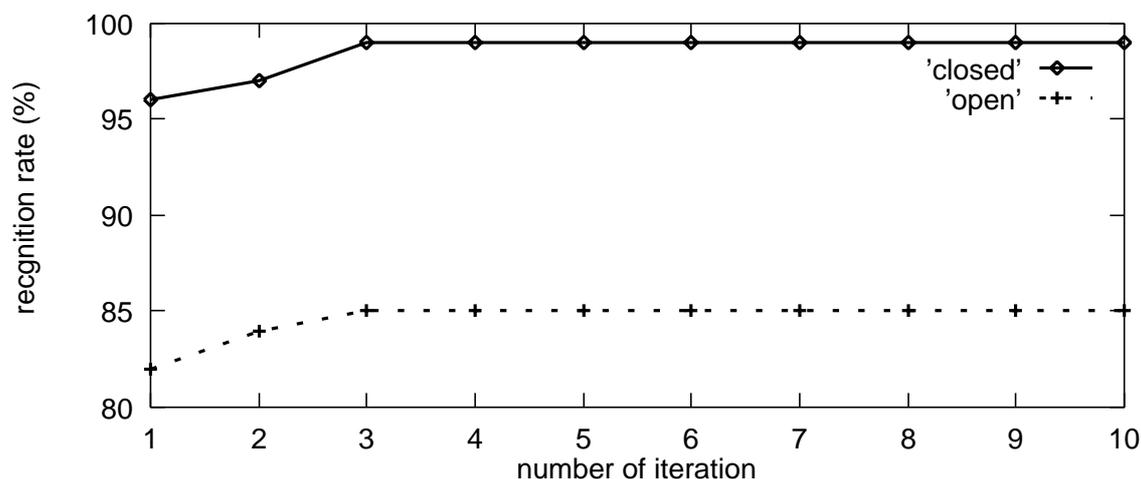
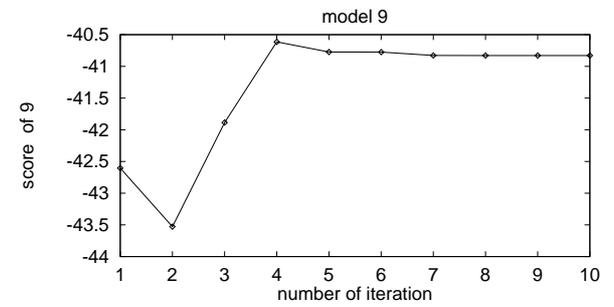
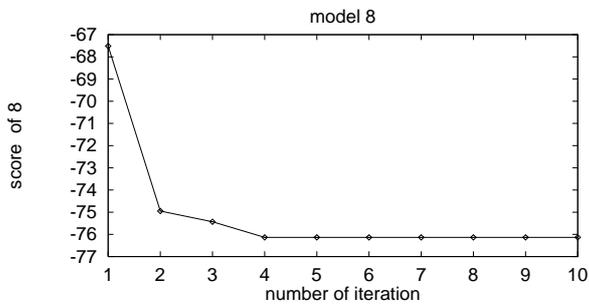
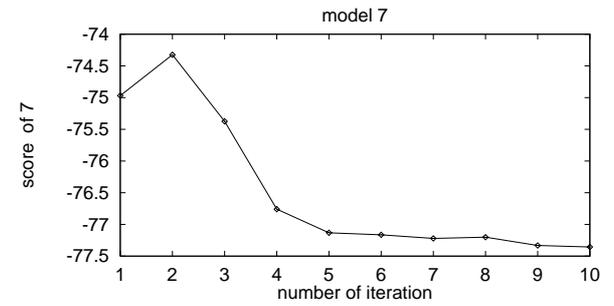
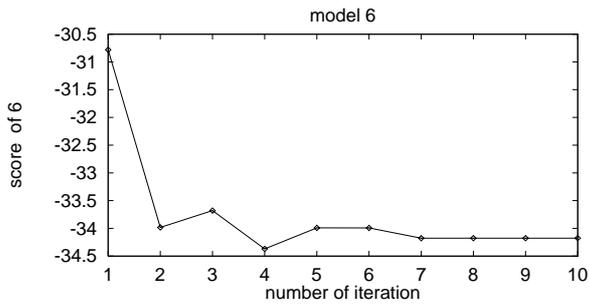
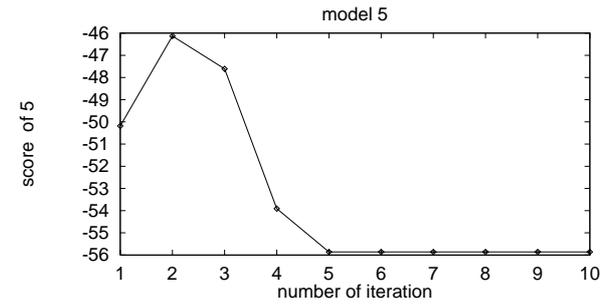
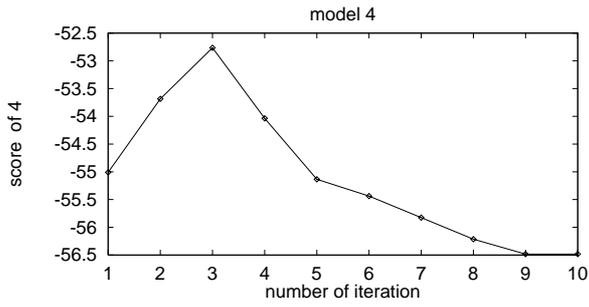
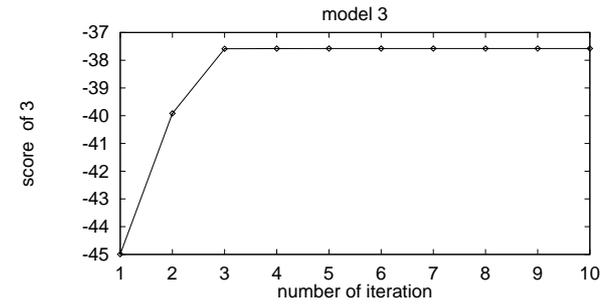
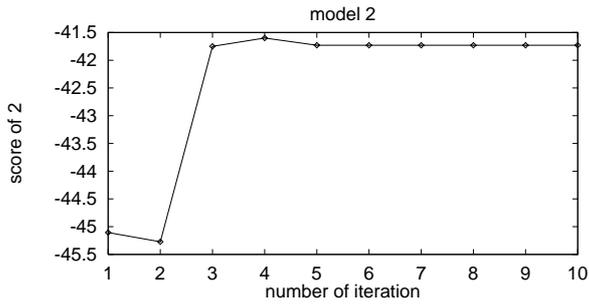
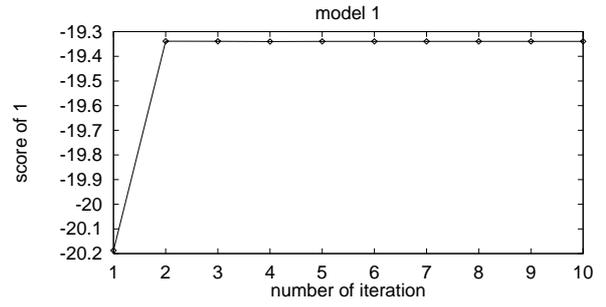
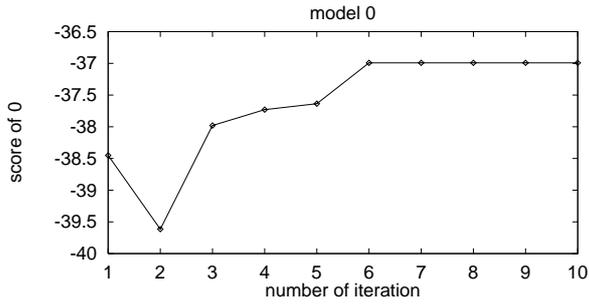


図 5.4: 距離によるクラス帰属度の認識率

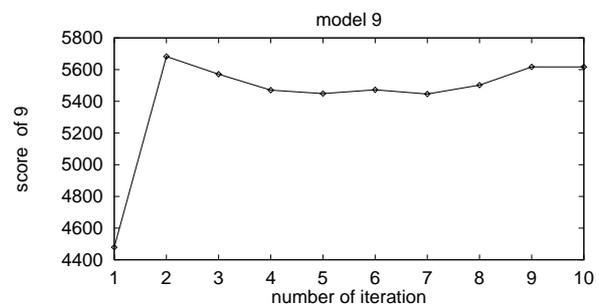
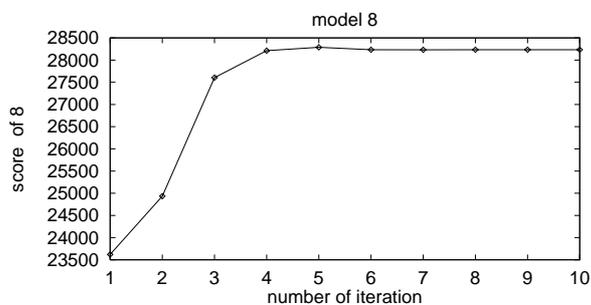
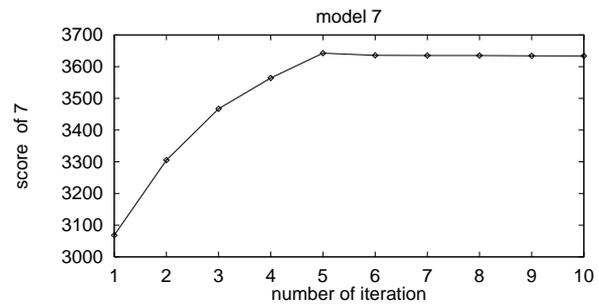
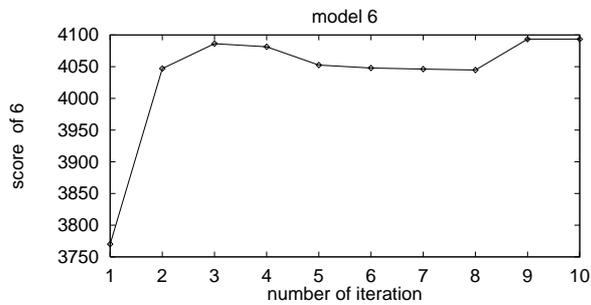
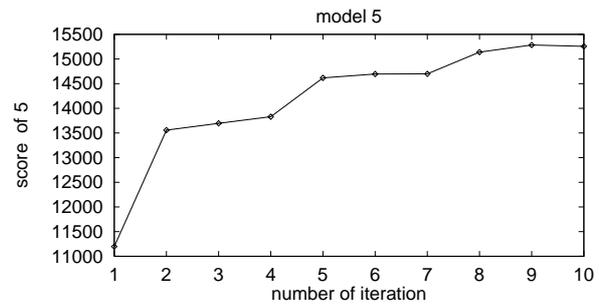
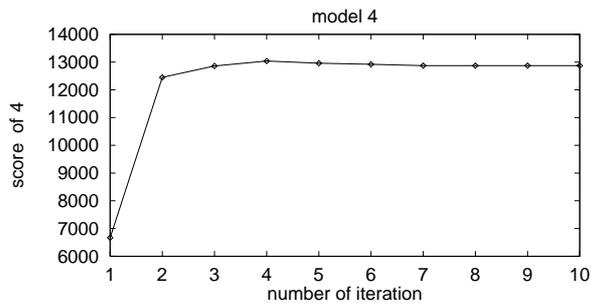
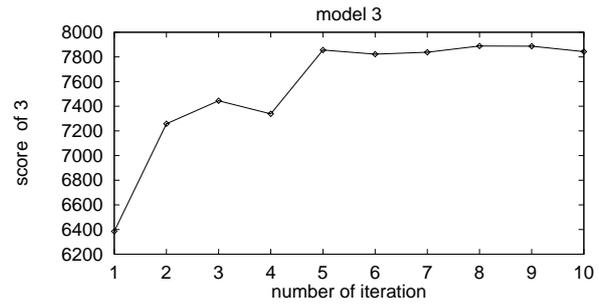
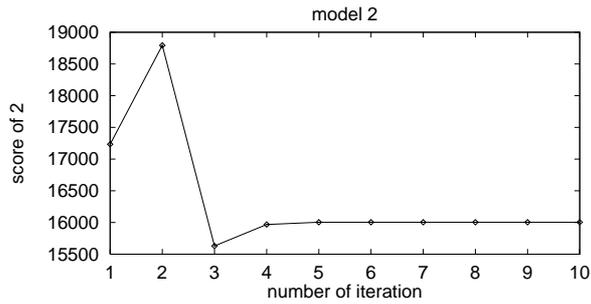
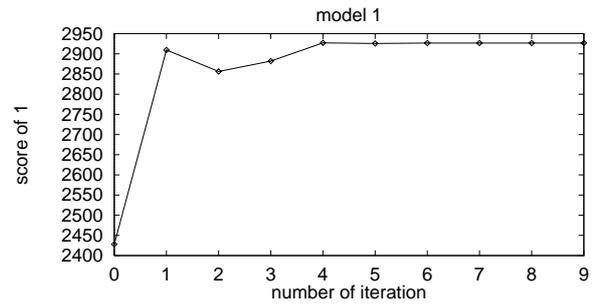
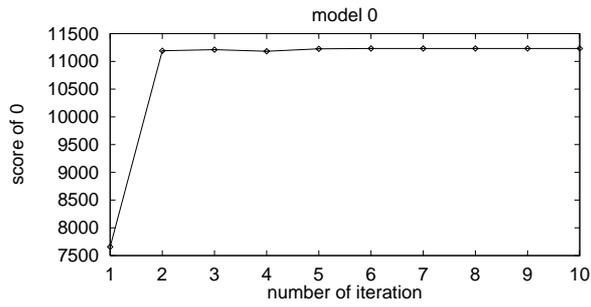
次項に 0~9 までの各モデル (model0 ~ model9) の反復学習によるスコアの変化を表したグラフを掲載した、グラフの横軸は反復学習の回数であり縦軸はスコアである、スコアは各モデルに認識対象になっているカテゴリの学習データを入力したときの出力の平均である。

実験 2 では認識率の点で比較すると実験 1 と同等の性能を示しているが、収束性が見られるカテゴリもあるが、全体としては収束しているとは言えない。



実験 2 では、単にスコアの変化だけを調査してもモデルの収束性が確認できなかった。次に各カテゴリの学習データを入力した場合に対応したカテゴリのモデルの出力するスコアと対応するカテゴリ以外のモデルから出力されたスコアで最も高いスコアの差について調査をした。各モデルのスコアの差を示したグラフを次項に掲載した、グラフの横軸は反復学習の回数であり縦軸はスコアの差である。

このグラフから model2 以外のモデルにおいて反復学習によりスコアの差が大きくなる傾向があり、この点において収束性があると言える。



5.5.3 実験 3

実験 3 では、学習の過程で各 SVC の目的クラスのデータについて識別関数の平均値を求め、その平均値と入力データの識別関数の差をクラス帰属度とする次式を用いた。このクラス帰属度はデータの分布をある程度考慮したものであり、前の 2 つの実験に用いたクラス帰属度と比較すると確率的な性質が強いと考えられるため、他の 2 つのクラス帰属度よりも Viterbi アルゴリズムに適応すると考えられる。

$$Mean_s = \frac{1}{N} \cdot \sum_{t=1}^N g_s(\mathbf{x}_t) \quad Mean_{s:s} \text{ 番目の SVC の平均値} \quad (5.7)$$

$$d_s(\mathbf{x}_t) = \sqrt{(Mean_s - g_s(\mathbf{x}_t))^2} \quad (5.8)$$

$$F_s(\mathbf{x}_t) = \exp(-d_s(\mathbf{x}_t)) \quad (5.9)$$

このクラス帰属度による認識率は学習用データに対して 100% であり、評価用データに対しては 84.4% であった。次のグラフに反復学習による認識率の変化を示す。

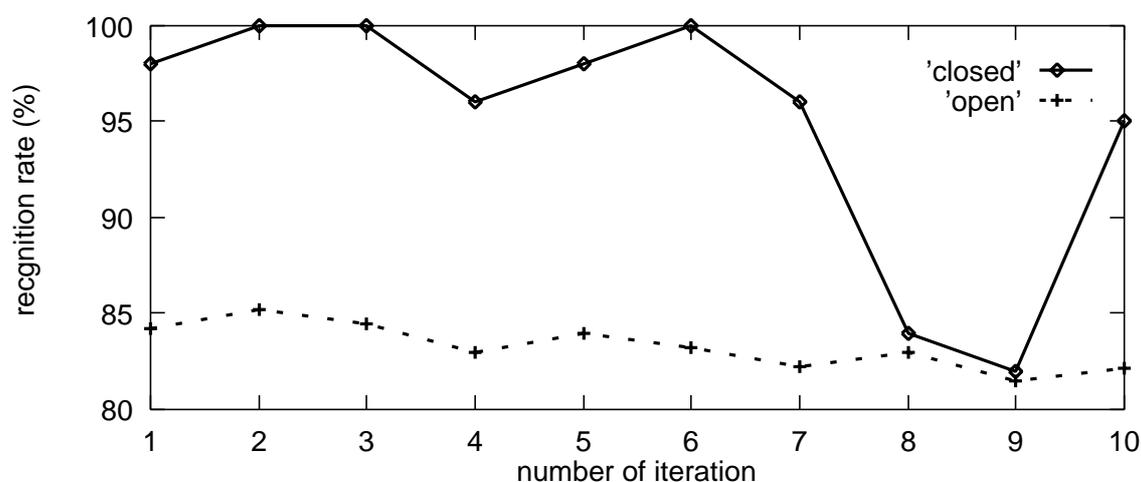
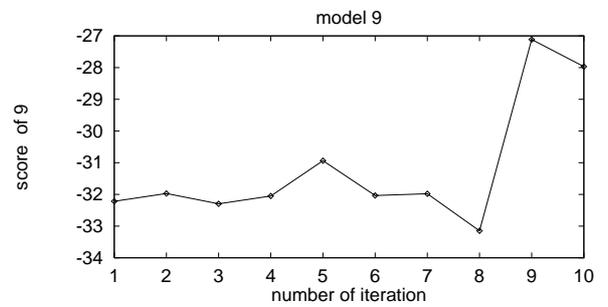
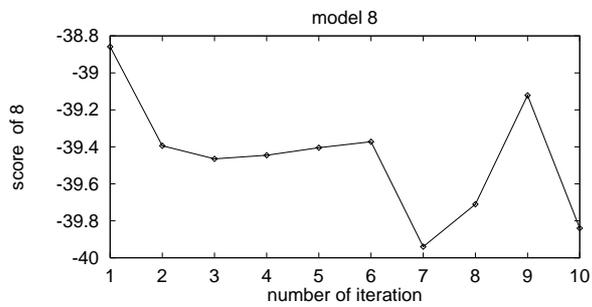
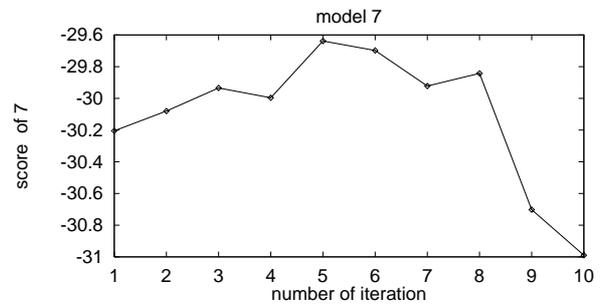
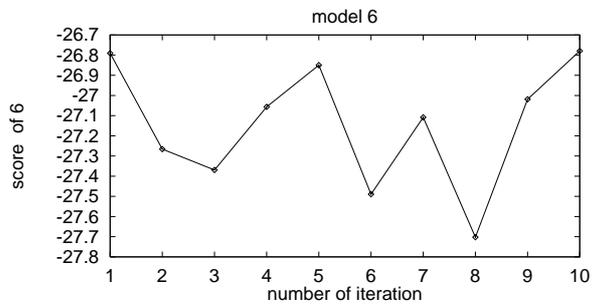
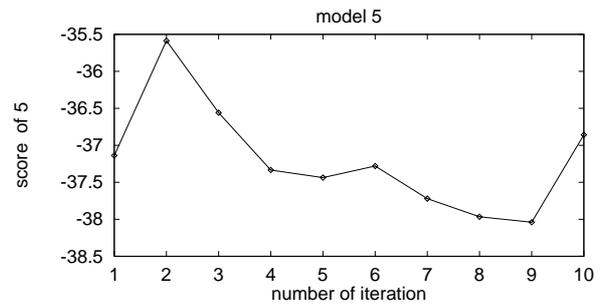
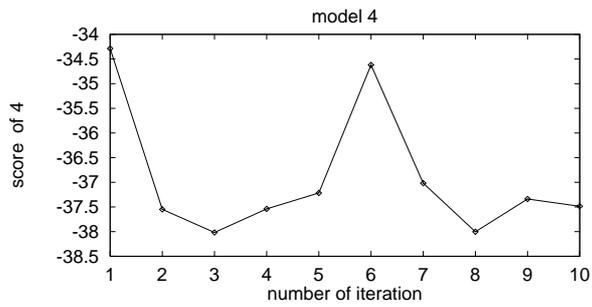
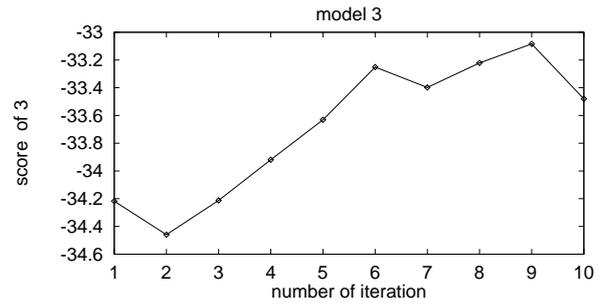
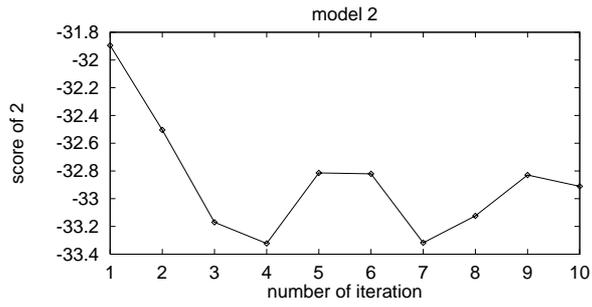
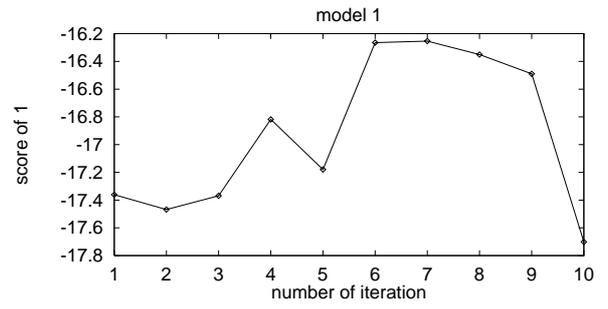
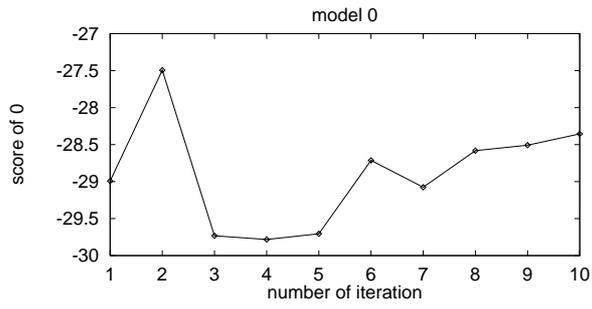


図 5.5: 距離によるクラス帰属度の認識率

次項に 0 ~ 9 までの各モデル (model0 ~ model9) の反復学習によるスコアの変化を表したグラフを掲載した、グラフの横軸は反復学習の回数であり縦軸はスコアである、スコアは各モデルに認識対象になっているカテゴリの学習データを入力したときの出力の平均である。

このグラフから反復学習によるスコアの収束性は確認されなかった、この原因としてはデータの分布が平均値だけでは十分に表現できなかった事と、その平均値が識別関数の値

による1次元の平均値であり、これも分布の表現を行なうには十分でなかったと考えられる。また、実験2と同様にスコアの差についても調査したが実験3においてはスコア差の収束性は見られなかった。



第 6 章

結論

6.1 研究結果

本論文では時系列データに対する新しい認識法として、SVC による時系列データの認識を試みた。SVC は識別境界を最適解として求める点や、経験的誤りと汎化能力の 2 点について最適化を行なうため未知のデータに対しても優れた汎化性能を示す点に特徴がある。これらの特徴により SVC は固定長データについては高い認識性能を示している。しかし SVC は認識対象となるデータの次元数が固定されていること仮定した識別器であるため、系列長の変化する時系列データを扱うことができなかった。そこで HMM が時系列データである音声信号に対して高い認識性能を示している点と時系列データのフレーム特徴量が固定長データであることに注目し、HMM と同じ構造を持つ CSVC を提案した。CSVC は固定長データを扱う SVC が複数個連結したモデルであるため、従来の SVC の定式化に大幅な変更を加える必要がなく、HMM で用いられるアルゴリズムを利用できる可能性がある。

CSVC は SVC の構成する最適分離超平面とデータ間の距離が、目的クラスからデータが生成される確からしさと考えここからクラス帰属度を定義し、また SVC 間の遷移の容易さを表すパラメータとして遷移の重みを導入しこれは遷移の回数を数えることで定義した。CSVC の認識のアルゴリズムは、遷移の重みとクラス帰属度の積を時系列データ全体についてさらに積をとりこれをスコアとし、最大のスコアをもつ遷移系列を探索する問題と定め、この問題を Viterbi アルゴリズムと同様のアルゴリズムにより求めるよう定義した。学習のアルゴリズムは、SVC が定式化において最小化問題の評価量を、連結した SVC 全てについて和を求め、和の最小化問題として CSVC の定式化を行なった、この最適化問題の解法として Viterbi 学習と類似の反復学習によるパラメータ推定を提案した。

この学習アルゴリズムの収束性については特定の条件が成立する場合に限り証明できる。

CSVC の認識性能を検証するため手書き数字の認識実験を行なった。時系列データのフレーム特徴量に 2 次元の速度ベクトルを用いた認識実験では HMM を下回る認識率であった、これは扱った特徴量にクラス間のオーバーラップが大きいというに特徴量の次元数が 2 次元と低かったためだと考えられる。フレーム特徴量を速度ベクトルと相対座標ベクトルの 4 次元とした認識実験では HMM と同等の認識率を示した。しかし特徴量 4 次元の場合でも学習の定式化で最小化問題とした評価関数の最小化は行なわれていなかった、この原因としては連結した個々の SVC の学習基準と CSVC が学習データの分割を決定する基準の間に大きなずれが存在すると考えられる。従って反復学習により学習データの最適な分割の決定ができていないと推測される。また、理論的な証明は行なっていないが CSVC のスコアによる評価を行なった。その結果、クラス帰属度として SVC の識別関数を用いた CSVC では 10 カテゴリ中 8 カテゴリでスコアの収束性が見られた、クラス帰属度に識別境界からの距離を用いた CSVC ではスコア自体の収束性は見られなかったがスコアの差においては 10 カテゴリ中 9 カテゴリで収束性が見られた。

提案した CSVC は学習アルゴリズムが適切でないにも関わらず HMM と同等に認識性能を示した、このため適切な学習アルゴリズムの導入により CSVC の認識性能がさらに向上する可能性がある。

6.2 今後の課題

今後の課題としては、CSVC の学習アルゴリズムを見直し、収束性のある定式化を行なうことである。また SVM による確率密度推定 [12, 13] の導入によるアルゴリズム、定式化についても検討する。

更に、音声認識などの問題に適応するには計算時間の削減 [14, 15] が必要になると考えられる。

謝辞

本研究を行うにあたり、有益な御指導と御助言を頂いた北陸先端科学技術大学院大学の嵯峨山茂樹教授に心から感謝致します。研究の方針や進め方について重要な御指導と御助言を頂きました同大学の下平博助教授に深く感謝致します。研究室の計算機環境を管理し、研究環境を整え、研究を進めるにあたり御助言を頂きました中井満助手に感謝致します。また手書き文字データベースを提供して頂いた東京工業江東専門学校 情報工学科の市村洋教授ならびに鈴木雅人に感謝します。最後に日頃から研究に関して御討論、御意見を頂きました嵯峨山・下平研究室の博士課程後期・前期在籍の学生ならびに研究生の皆様に感謝致します。

参考文献

- [1] Christopher J.C. Burges : “A Tutorial on Support Vector Machines for Pattern Recognititon,”
a draft to Data Mining and Knowlledge Discovery 19/Nov/1996.
- [2] Corinna Cortes, Vladimir Vapnik, “Support-Vector Networks”
Machine Learning, 20, 273-297, 1995.
- [3] 坂井護, “Support Vector Machine を用いたパターン認識に関する研究”
北陸先端科学技術大学院大学 情報科学研究科 修士論文, 1999
- [4] M.O.Stitson, J.A.E.Weston, A.Gammerman, V.Vovk, V.Vapnik, “Theory of Support
Vector Machines”
Technical Report, Royal Holloway University o London, April 1998.
- [5] <http://www.clopinet.com/isabelle/Projects/SVM/applist.html>
- [6] 磯 健一 渡辺 隆夫 : “ニューラル予測モデルを用いた不特定話者音声認識,”
信学論 D-II, Vol.J73-D-II, No. 8, pp.1315-1321, August 1990
- [7] 加藤喜永、杉山雅英. “ファジィパーテーションモデルを用いた連続音声認識”
信学技法, SP92-28, 1992.
- [8] Philip Clarkson : “On The Use of Support Vector Machines for Phonetic Classifica-
tion,” InICASSP99, pp.585-588, 1999.
- [9] Partha Niyogi, Chris Burges, Padma Ramesh “Distinctive Feature Detection Using
Support Vector Machines”
IEEE1999, pp.425-428, 1999.

- [10] Vladimir N. Vapnik. “Statistical Learning Theory”
Wiley.
- [11] Vladimir Vapnik, Steven E. Golowich, Alex Smola, “Support Vector Method for
Function Approximation Regression Estimation, and Signal Processing”
Neural Information Processing Systems, Vol9 MIT Press, 1997.
- [12] Jason Weston, Alex Gammerman, Mark O. Stitson, Vladimir Vapnik, Volodya Vovk,
Chris Watkins : “Support Vector Density Estimation,”
Advanced Kernel Methods, pp.293-306. MIT Press, 1998.
- [13] Bernhard Scholkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, “Estimating
the Support of a High-dimensional Distribution”
<http://svm.first.gmd.de/>
- [14] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, “Improvements to
Platt’s SMO Algorithm for SVM Classifier Design”
Tech Report, <http://svm.first.gmd.de/>
- [15] 津田宏治, “ヒルベルト空間における部分空間法”
信学技法, NC98-36, pp47-54, 1998.
- [16] Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of Speech Recognition”
Prentice-Hall.
- [17] 中川聖一, “確率モデルによる音声認識”
電子情報通信学会.
- [18] 鳥脇純一郎, “認識工学 パターン認識とその応用”
コロナ社.
- [19] 田代裕貴, 市村洋, オンライン手書き教育漢字データベースの作成.
情報処理学会第 52 回全国大会, 第 2 巻, pp.249-250, 1996.

研究業績一覧

- 中井浩一, 中井満, 下平博, 嵯峨山茂樹, “Support Vector Machine による時系列パターンの認識 “
電子情報通信学会, 信学技法, PRMU99, pp15-20, 1999, 愛媛大学工学部.