TR－A－0123

識別学習理論による音声認識

片桐　滋

1991.11. 1
（1991.10.25受付）

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷　☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani, Sanpeidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

識別学習理論による音声認識

ＡＴＲ視聴覚機構研究所
聴覚研究室

片桐　滋

概要

本技術覚書は、我々が数年来研究を進めてきた２つの概念、１）一般化確率的降下法（ＧＰＤ：Generalized Probabilistic Descent Method）と２）これを最も実際的な実現法として用いる最小分類誤り学習理論（ＭＣＥ：Learning for Minimum Classification Error）、に関する要旨を講演用のスライド形式でまとめたものである。標題にある音声認識は我々の研究分野ではあるが、本覚書で紹介される概念は、音声認識のみならず多様なパタン認識に用いることができる。

人工神経回路網や学習ベクトル量子化の到来によって、古典的識別学習が再び注目を集めるようになった。しかし、魅力あふれるこれらの新技術においてさえも、実際に行われている識別学習は極めて直観的である。実際、その多くは分類器構造固有の特殊な学習手続きとして経験的に構築されたものであり、学習結果の最適性等の学習性質もほとんど解析されていない。多様なパタン認識の応用に耐え得る普遍的識別学習法はいまだ存在しないと言っても過言ではないだろう。ＧＰＤは、こうした状況を解決するために、四半世紀も以前に開発された確率的降下法を現代向きに改編したものである。一般化の主たる貢献は、確率的降下法の数学的不備を補うことによって勾配法を用いるのに適した厳密な定式化を与えたことと、音声信号のような可変長パタンの分類を可能にする多種多様な分類器構造のための一群の新しい識別学習法を実現したことである。また、この研究の中で、学習ベクトル量子化のような従来の直観的学習アルゴリズムに理論的背景を与えたことや一般的多層ネットワークを用いた新しい識別ネットワークを開発したこともここで付言すべきであろう。

ＧＰＤの出現によって、学習手順に関する実際的問題は著しく改善される。パタン認識の究極の目的が誤分類の最小化、言い替えればベイズの最小誤分類確率の実現にあることは言うまでもない。しかし、新しい人工神経回路網の研究においてさえも最も一般的に行なわれているパーセプトロン損失価値や最小自乗誤差損失価値を用いた識別学習は、明らかにこの本来の目的を追う道筋から逸脱している。ＭＣＥは、こうした現状に対する極めて新しい研究方向を示している。即ちここでは、最も現実的かつ効果的な最適探索法である勾配法によって最小分類誤りの直接の近似を行えることが明確に示されている。

上述の２つの概念の、理論的のみならず実際的な卓越性が一連の実験によって示されている。実験には、有名なフィッシャーの水仙分類課題や極めて認識困難な類似音節の分類課題を用いている。実験結果は、いずれも、紹介された新概念の優れた有効性を示している。

スライド（講演）の流れ

１．概要　２．統計的パタン分類　３．動機と着想　４．最小自乗誤差損失価値を用いた識別学習　－多層パーセプトロンの場合－　５．複数参照ベクトルから成る距離分類器のためのＬＶＱ　－ＬＶＱ２の場合－　６．確率的降下法　７．確率的降下法の収束定理　８．一般化確率的降下法　９．判別関数の例　１０．誤分類測度の例　１１．損失価値関数　１２．最小分類誤りの勾配法最適探索に適した新しい定式化　１３．識別学習アルゴリズムの例（複数参照ベクトルを持つ距離分類器の場合；ＬＶＱの理論的背景）　１４．一般化多層前向きネットワーク　１５．ＬＶＱと３層識別ネットワークとの関連　１６．実験例　１７．まとめ

参考文献

1. Shigeru KATAGIRI; Systematic Explanation of Learning Vector Quantization and Multi-Layer Perceptron: Proposition of Distance Network, IEICE, Tech. Report MBE88-72, pp.75-82 (1988. 10).

2. Shigeru KATAGIRI, Chin-Hui LEE, and Biing-Hwang JUANG; A Generalized Probabilistic Descent Method, ASJ, Proc. of Fall Meeting, 2-P-6, pp.141-142 (1990. 9).

3. Takashi KOMORI and Shigeru KATAGIRI; A New Discriminative Training Algorithm for Dynamic Time Warping-Based Speech Recognition, IEICE, Tech. Report SP91-10, pp.33-40 (1991. 6).

4. Erik MCDERMOTT and Shigeru KATAGIRI; Discriminative Training for Various Speech Units, IEICE, Tech. Report SP91-12, pp.47-54 (1991. 6).

5. Shigeru KATAGIRI, Chin-Hui LEE, and Biing-Hwang JUANG; Discriminative Multi-Layer Feed-Forward Networks, IEEE, Neural Networks for Signal Processing - Proc. of the 1991 IEEE-SP Workshop (1991. 9).

6. Shigeru KATAGIRI, Chin-Hui LEE, and Biing-Hwang JUANG; New Discriminative Training Algorithms Based on the Generalized Probabilistic Descent Method, IEEE, Neural Networks for Signal Processing - Proc. of the 1991 IEEE-SP Workshop (1991. 9).

7. Takashi KOMOTI, and Shigeru KATAGIRI; A Discriminative Training for DTW-Based Pattern Recognition, ASJ, Proc. of Fall Meeting (1991. 10).

8. Erik MCDERMOTT and Shigeru KATAGIRI; Prototype Based Minimum Error Classification for Various Speech Units, ASJ, Proc. of Fall Meeting (1991. 10).

9. Biing-Hwang JUANG and Shigeru KATAGIRI; Discriminative Learning for Minimum Error Classification, under submission.

謝辞

本覚書に紹介した内容は、Biing-Hwang Juang博士とChin-Hui Lee博士との共同研究の成果である。本覚書が内部資料であるために両氏を著者欄から除いたことを詫びるとともに、実に多くの貢献を得たことを明記したい。また、Wu Chou博士とBiing-Hwang Juang博士のＧＰＤの学習収束性に関する研究は多くの示唆を与えてくれた。Pao-Chung Chang博士はＧＰＤを用いた音声認識実験を遂行してくれた。Erik McDermottと小森　隆、Andrew Duchon のＡＴＲにおける共同研究者３氏には、実り多い討論と多くの助力を得た。ここで数え上げることができない程多くのＡＴ＆Ｔベル研究所とＡＴＲの研究者諸氏にも実に多くの助言を頂いた。これら全ての方々に深く感謝する次第である。

識別学習理論による音声認識

ATR視聴覚機構研究所
聴覚研究室

片桐　滋

概要

- 統計的パタン認識
  - ベイズ決定則
    - 1）最尤法
    - 2）最大化ー期待値法
  - 判別関数法

- 動機
  - 隠れマルコフモデル(HMM)
  - 人工神経回路網(ANN)
  - 学習ベクトル量子化(LVQ)
  - 高精度分類能力
  - 動的（可変長）パタン分類

概要（続き）

- 主たる概念
  - ◆ 一般化確率的降下法(GPD)
  - ◆ 最小分類誤り学習(MCE)
  - ◆ 一般化多層前向きネットワーク (FNN)

- 貢献
  - ◆ 勾配法のためのＭＣＥ理論
  - ◆ 種々の識別学習法
    - １） 動的パタン識別法
    - ２） 識別FNN
    - ３） その他

統計的パタン分類

● ベイズ決定則

$$C(\mathbf{x}) = C_i \qquad \text{if } p_\Lambda(C_i|\mathbf{x}) = \max_j \left\{ p_\Lambda\left(C_j|\mathbf{x}\right) \right\}$$

又は

$$C(\mathbf{x}) = C_i \ \text{ if } p_\Lambda\left(\mathbf{x}|C_i\right)p_\Lambda(C_i) = \max_j \left\{ p_\Lambda\left(\mathbf{x}|C_j\right)p_\Lambda\left(C_j\right) \right\}$$

● ベイズ決定法

- ◆ $p_\Lambda(C_i|\mathbf{x})$の推定
- ◆ $p_\Lambda(\mathbf{x}|C_i)$の推定
  - 1） 最尤法
  - 2） 最大化ー期待値法

統計的パタン分類（続き）

- 判別関数法
  - ◆ 判別関数　$g_i(\mathbf{x};\ \Lambda)$
  - ◆ 決定則

$$C(\mathbf{x}) = C_i \qquad \text{if } g_i(\mathbf{x};\ \Lambda) = \max_j \left\{ g_j(\mathbf{x};\ \Lambda) \right\}$$

  - ◆ 標本損失危険度の最小化
    - 1）パーセプトロン誤差
    - 2）最小自乗誤差(MSE)

- LVQ

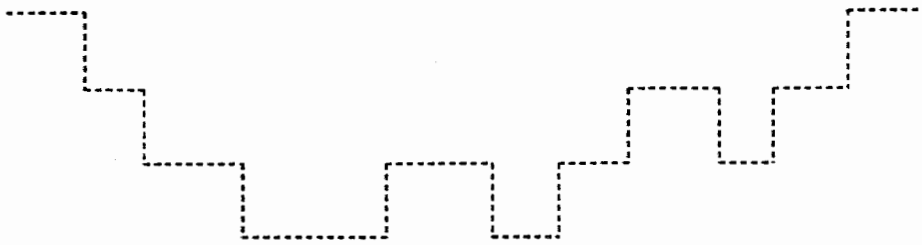
動機と着想

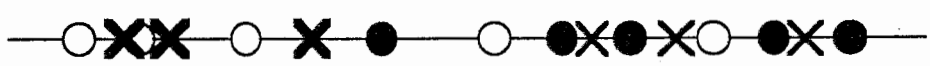
- 最尤法あるいはこれに準ずる方法に基づく慣習的分類器の学習
  - 距離分類器
  - 非線形時間軸伸縮を伴うパタン整合法(DTW)
  - HMM

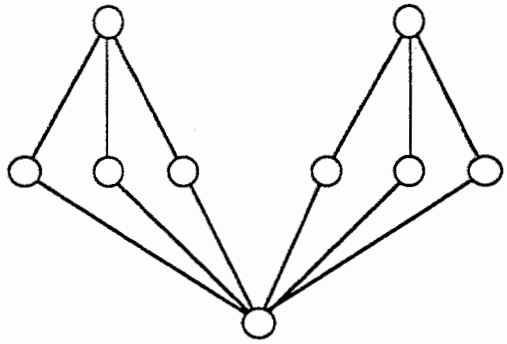
動機と着想（続き）

- 分類能力の向上を目指した従来策とその問題
  - ◆ 誤り訂正学習
  - ◆ 相互情報量の利用
  - ◆ MSE損失価値を用いた識別学習
  - ◆ LVQ
    1）経験則（最適性等の学習に関する性質が不明）
    2）最小分類誤りとの一貫性の欠如
  - ◆ 確率的降下法（PD）
    1）静的（固定次元）ベクトルしか扱えない
    2）定式化における数学的不備
- GPD
- MCE

MSE損失価値を用いた識別学習
－多層パーセプトロンの場合－







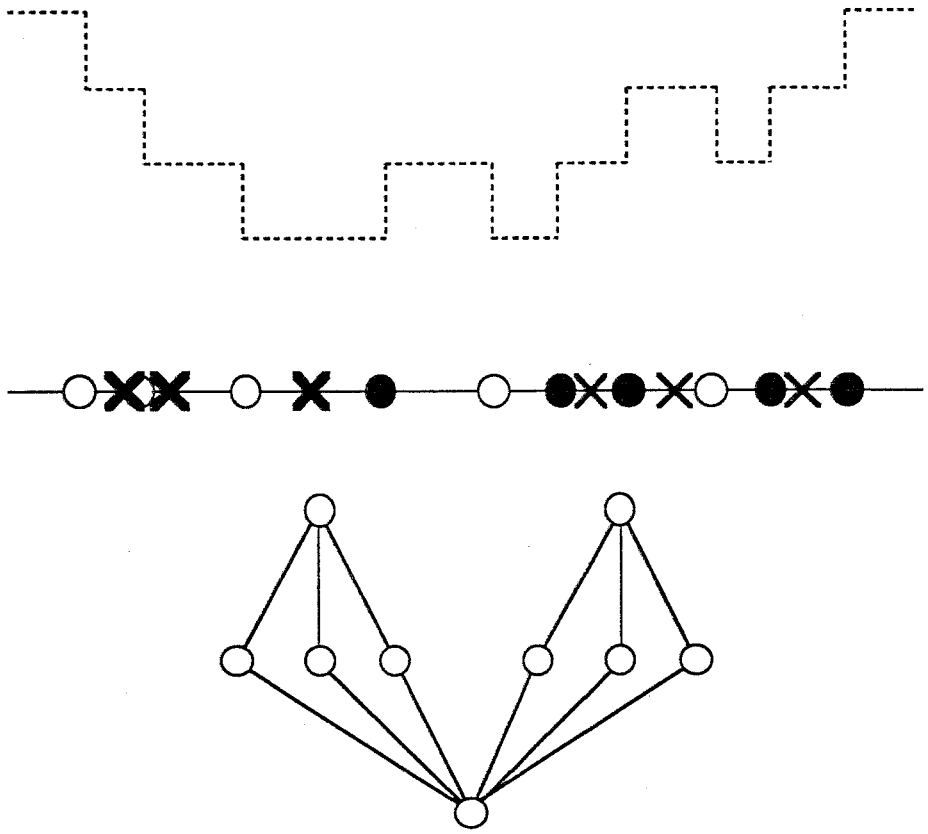
- 階段関数を用いたMSE学習；高精度分類

(1 0) (0 1)

- 無限個の標本を用いたMSE学習；事後確率の良い推定
- 決定則との一貫性の欠如

$$C(\mathbf{x}) = C_i \qquad \text{if } g_i(\mathbf{x};\ \Lambda) = \max_j \left\{ g_j(\mathbf{x};\ \Lambda) \right\}$$

複数参照ベクトルから成る
距離分類器のためのLVQ
－LVQ2の場合－



- 判別関数；ユークリッド距離
- 損失価値；直観的最小分類誤り
- 修正規則；直観的修正則

確率的降下法(PD)

- 貢献
  - 分類要因を関数形式に埋め込む
  - 学習収束の性質を明かにする
- 要点
  - 判別関数　　$g_i(\mathbf{x};\ \Lambda)$
    1）線形判別関数
    2）区間線形判別関数
  - 誤分類測度

$$d_k(\mathbf{x})$$

  - 損失価値

$$\ell_k(\mathbf{x};\ \Lambda) = \ell\big(d_k(\mathbf{x})\big)$$

  - 期待損失価値

$$L(\Lambda) = \sum_k \int \ell_k(\mathbf{x};\ \Lambda) p\big(\mathbf{x},\ C_k\big) d\mathbf{x}$$

  - 収束定理
- 問題点
  - 静的（固定次元）パタンのみを対象とする
  - 不連続等の数学的不備
  - 実装（応用）可能性が小さい

PDの収束定理

与えられた標本に関し$\mathbf{x} \in C_k$と仮定する。もし、分類器のパラメタの修正量$\delta\Lambda(\mathbf{x}, C_k, \Lambda)$が以下のように設定されるものとすれば、

$$\delta\Lambda(\mathbf{x}, C_k, \Lambda) = -\varepsilon \mathbf{U} \nabla \ell_k(\mathbf{x}; \Lambda) \quad (1)$$

（但しここで、$\mathbf{U}$は正定値行列でεは小さな正の実数）

そのとき

$$E[\delta L(\Lambda)] \leq 0 \quad (2)$$

が成り立つ。さらに、もしランダムな標本の無限個列$\mathbf{x}_i$が学習（訓練）に用いられ、かつ（1）式の修正規則が以下の条件を満たすステップサイズ列ε_iとともに用いられるものとすれば、

$$\sum_{i=1}^{\infty} \varepsilon_i \to \infty \quad (3)$$

$$\sum_{i=1}^{\infty} \varepsilon_i^2 < \infty \quad (4)$$

そのとき（5）式に基づくパラメタ列Λ_iはの局所的最小をもたらすΛ^*に確率１で収束する。

$$\Lambda_{i+1} = \Lambda_i + \delta\Lambda(\mathbf{x}_i, C_k, \Lambda_i) \quad (5)$$

一般化確率的降下法(GPD)

- 要点
  - 動的パタンを扱うための確率測度の導入
  - 一般化判別関数の定義
  - 一般化誤分類測度の定義
    Lpノルム形式の利用
  - 種々の連続損失価値の導入
    特にシグモイド形式
  - 経験的平均損失価値の利用
  - 確率等の制約を考慮した収束定理の証明

GPD（続き）

- 貢献
  - 動的パタンを扱うことができる
  - 高い一般性

    Lpノルム形式の採用に起因する
  - 勾配法に適した厳密な定式化
  - 最小分類誤り問題の新しい定式化
  - 具体的な種々の識別学習アルゴリズムの開発
  - 従来の学習法（例:LVQ）に理論的背景を与える
  - 分類器の設計のみならず特徴抽出やセグメンテーション等を統一的に扱う新しい枠組を提供

判別関数の例

- 距離

$$g_j(\mathbf{x};\Lambda) = D_\Lambda\left(\mathbf{x}\middle|C_j\right)$$

- 確率

$$g_j(\mathbf{x};\Lambda) = p_\Lambda\left(\mathbf{x}\middle|C_j\right)$$

- 尤度

$$g_j(\mathbf{x};\Lambda) = \ln\left\{p_\Lambda\left(\mathbf{x}\middle|C_j\right)\right\}$$

- 相互情報量

$$g_j(\mathbf{x};\Lambda) = \ln\left\{\frac{p_\Lambda\left(\mathbf{x}\middle|C_j\right)}{\sum_i p_\Lambda\left(\mathbf{x}\middle|C_i\right)}\right\}$$

- 確率比

$$g_j(\mathbf{x};\Lambda) = 1 - \frac{p_\Lambda\left(\mathbf{x}\middle|C_j\right)}{\sum_i p_\Lambda\left(\mathbf{x}\middle|C_i\right)}$$

- 一般化距離

$$g_j(\mathbf{x};\Lambda) = \left[\sum_{b=1}^{B_m}\left\{D\left(\mathbf{x}_1^T,\mathbf{r}_j^b\right)\right\}^{-\zeta}\right]^{-1/\zeta}$$

$$\text{ここで} D\left(\mathbf{x}_1^T,\mathbf{r}_j^b\right) = \left[\sum_{\theta=1}^{\Theta}\left\{D_\theta\left(\mathbf{x}_1^T,\mathbf{r}_j^b\right)\right\}^{-\psi}\right]^{-1/\psi}$$

- 一般化尤度

$$g_j\left(\mathbf{x}_1^T;\Lambda\right) = \ln\left[\sum_{\theta=1}^{\Theta_j}\left\{p_\Lambda\left(\mathbf{x}_1^T,\theta\middle|C_j\right)\right\}^{\eta}\right]^{1/\eta}$$

誤分類測度の例

$$\mathbf{x} \in C_k$$

- 距離に基づく場合

$$d_k\left(\mathbf{x}_1^T\right) = g_k\left(\mathbf{x}_1^T;\Lambda\right) - \left[\frac{1}{M-1}\sum_{j,j\neq k}\left\{g_j\left(\mathbf{x}_1^T;\Lambda\right)\right\}^{-\zeta}\right]^{-1/\zeta}$$

- 尤度に基づく場合

$$d_k\left(\mathbf{x}_1^T\right) = -g_k\left(\mathbf{x}_1^T;\Lambda\right) + \ln\left[\frac{1}{M-1}\sum_{j,j\neq k}\exp\left\{\zeta g_j\left(\mathbf{x}_1^T;\Lambda\right)\right\}\right]^{1/\zeta}$$

損失価値関数

($\mathbf{x}_1^T \in C_k$　の場合)

● 指数（甘利）

$$\ell_k\left(\mathbf{x}_1^T;\Lambda\right)=\begin{cases}(d_k)^{\sigma} & d_k>0\\ 0 & d_k\le 0\end{cases}$$

● シグモイド

$$\ell_k\left(\mathbf{x}_1^T;\Lambda\right)=\frac{1}{1+e^{-\alpha(d_k+\beta)}},\quad \alpha>0$$

最小分類誤りの勾配法最適探索に適した新しい定式化

● ベイズ最小危険度（最小分類誤り）

$$\mathcal{E} = \sum_{k=1}^{M} \int_{\mathcal{X}_k} p_\Lambda\left(\mathbf{x}_1^T, C_k\right) 1\left(\mathbf{x}_1^T \in C_k\right) d\mathbf{x}_1^T$$

ここで、

指示関数

$$1(\mathcal{A}) = \begin{cases} 1, & \text{if } \mathcal{A} \text{ is true} \\ 0, & \text{otherwise} \end{cases}$$

$$\mathcal{X}_k = \left\{ \mathbf{x}_1^T \in \mathcal{X} \middle| p_\Lambda\left(C_k \middle| \mathbf{x}_1^T\right) \neq \max_i p_\Lambda\left(C_i \middle| \mathbf{x}_1^T\right) \right\}$$

$$\begin{aligned} &\mathcal{E} \\ &= \sum_{k=1}^{M} \int_{\mathcal{X}} p_\Lambda\left(\mathbf{x}_1^T, C_k\right) 1\left(\mathbf{x}_1^T \in C_k\right) \times \\ &\qquad 1\left(p_\Lambda\left(C_k \middle| \mathbf{x}_1^T\right) \neq \max_i p_\Lambda\left(C_i \middle| \mathbf{x}_1^T\right) \right) d\mathbf{x}_1^T \\ &\approx \sum_{k=1}^{M} \int_{\mathcal{X}} p_\Lambda\left(\mathbf{x}_1^T, C_k\right) 1\left(\mathbf{x}_1^T \in C_k\right) \ell\left(d_k\left(\mathbf{x}_1^T\right)\right) d\mathbf{x}_1^T \end{aligned}$$

識別学習アルゴリズムの例
(複数参照ベクトルを持つ距離分類器の場合；LVQの理論的背景)

- 判別関数

$$g_j(\mathbf{x};\Lambda)=\left[\sum_{i=1}^{I_j}\left\{D\left(\mathbf{x};\lambda_{ji}\right)\right\}^{-\gamma}\right]^{-1/\gamma}$$

ここで$D\left(\mathbf{x};\lambda_{ji}\right)=\left(\mathbf{x}-\mathbf{r}_{ji}\right)^t\left(\Sigma_{ji}\right)^{-1}\left(\mathbf{x}-\mathbf{r}_{ji}\right)+\ln\left|\Sigma_{ji}\right|$

- 誤分類測度　$\mathbf{x}\in C_k$

$$d_k(\mathbf{x})=g_k(\mathbf{x};\Lambda)-\left[\frac{1}{M-1}\sum_{j,j\neq k}\left\{g_j(\mathbf{x};\Lambda)\right\}^{-\zeta}\right]^{-1/\zeta}$$

- 簡単化

$$\gamma,\zeta\to\infty,\qquad D\left(\mathbf{x};\lambda_{ji}\right)=\left\|\mathbf{x}-\mathbf{r}_{ji}\right\|^2$$

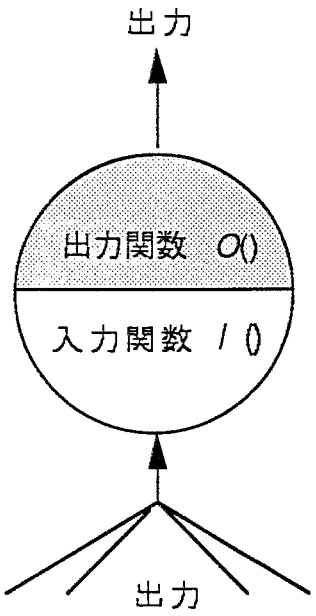
- 修正規則
  - クラスk

$$2\varepsilon\mathbf{U}\ell'_k\left(d_k(\mathbf{x})\right)\left(\mathbf{x}-\mathbf{r}_{kc_k}\right)$$

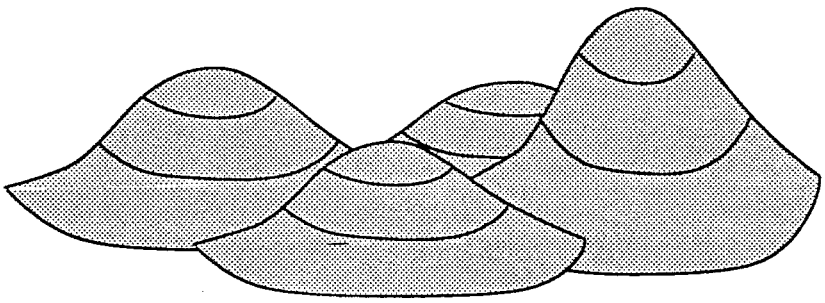
  - クラスβ（最大尤度しかしk以外のクラス）

$$-2\varepsilon\mathbf{U}\ell'_k\left(d_k(\mathbf{x})\right)\left(\mathbf{x}-\mathbf{r}_{\beta c_\beta}\right)$$

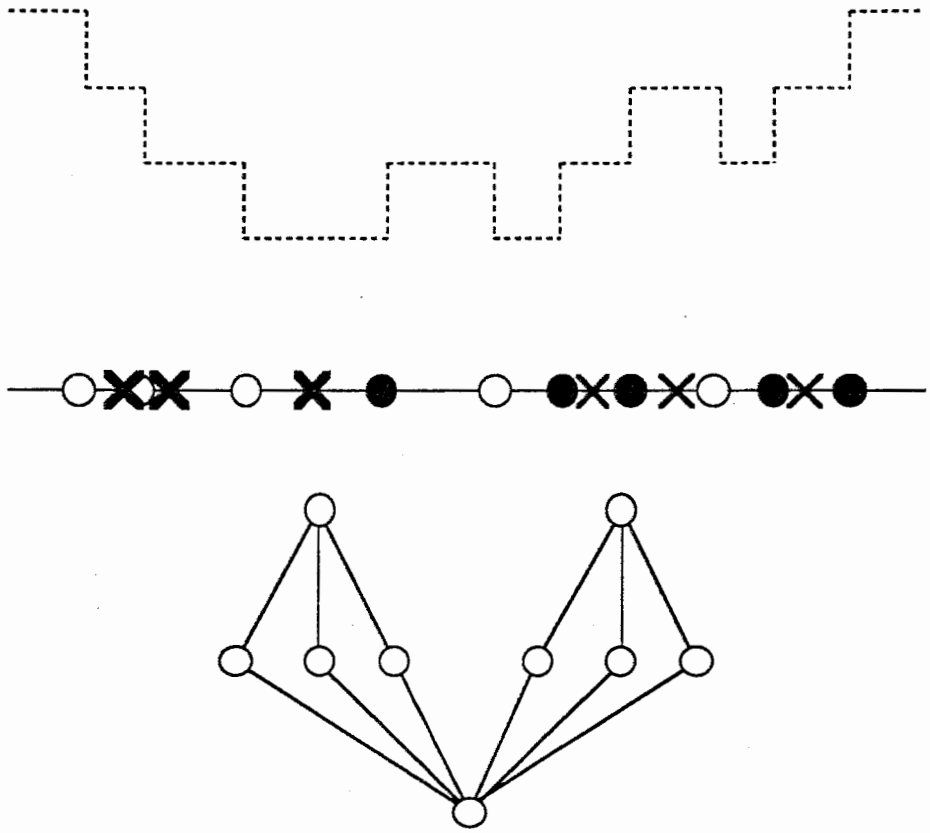
一般化多層前向きネットワーク



- 入力関数の選択
  - ◆ 内積；パーセプトロンネットワーク
  - ◆ 距離；距離ネットワーク
  - ◆ 尤度；尤度ネットワーク



LVQと３層識別距離ネットワークとの関連



- 最小分類誤りと一貫性のある経験的平均損失価値
- 勾配計算のチェーン規則
- 新しいクラス距離（判別関数）

$$i_{3j} = \left\{ \frac{1}{L_j} \sum_{i \in Q_j} (o_{2i})^{-\xi} \right\}^{-\frac{1}{\xi}}$$

  - ◆ 遠藤らによる修正LVQ2、ファジーLVQ
- $\xi, \eta \to \infty$
- 近似的損失価値
  - ◆ McDermottによる修正LVQ2

実験例

● フィッシャーの水仙分類課題

－4次元、3クラス－

◆ 線形分類器（パーセプトロン損失価値）

86.0%

◆ 3層パーセプトロン（MSE）

89.5%

◆ 3層パーセプトロン（MCE,GPD）

97.8%

◆ 3層距離ネットワーク（MCE,GPD）

98.7%

◆ 簡単化3層距離ネット（LVQ:ユークリッド距離）

98.7%

◆ 簡単化3層距離ネット（LVQ:無相関尤度距離）

100.0%

実験例（続き）

● Eセット英語音声分類課題

－可変次元、９クラス－

◆ DTW

従来型　－＞　GPD, MCE

１）１参照ベクトル／クラス

58.0%　　->　　*78.4%*

２）４参照ベクトル／クラス

63.8%　　->　　*83.4%*

３）１２参照ベクトル／クラス

67.6%

◆ HMM

従来型　－＞　GPD, MCE

１）５状態、５混合ガウス型

61.7%

２）１０状態、５混合ガウス型

66.7%

３）１５状態、５混合ガウス型

69.0%　　->　　*85.7%*

まとめ

- 一般化確率的降下法
- 最小分類誤り問題の新しい定式化
- 種々の分類器構造のための新しい識別学習アルゴリズム
- 従来のアルゴリズムの理論的背景
- 実験例