

IFAC

INTERNATIONAL FEDERATION
OF AUTOMATIC CONTROL



WARSZAWA 1969

Learning Systems and Pattern Recognition

Fourth Congress of the International
Federation of Automatic Control
Warszawa 16–21 June 1969

TECHNICAL
SESSION

21



Organized by
Naczelna Organizacja Techniczna w Polsce

INTERNATIONAL FEDERATION OF AUTOMATIC CONTROL

Learning Systems and Pattern Recognition

TECHNICAL SESSION No 21

FOURTH CONGRESS OF THE INTERNATIONAL
FEDERATION OF AUTOMATIC CONTROL
WARSZAWA 16 – 21 JUNE 1969



Organized by
Naczelna Organizacja Techniczna w Polsce



Biblioteka
Politechniki Białostockiej



1181045

K-1294

Contents

Paper No			Page
21.1	SU	- Ya. Z. Tsypkin, G. K. Kelmans, L. E. Epstein - Learning Control Systems.....	3
21.2	USA	- M. H. Hammond, H. R. Weed - Determination of the Optimal Measurement Sequence for a Noisy Pattern Recognition System via the Discrete Maximum Principle	21
21.3	SU	- E. M. Braverman, B. M. Litvakov - Convergence of Learning and Adaptive Algorithms.....	34
21.4	USA	- K. Fukunaga, T. F. Krile - An Algorithm for Calculating Recognition Errors in Pattern Recognition....	49
21.5	PL	- Z. Bubnicki - Least Interval Pattern Recognition and its Application to Control Systems.....	62
21.6	H	- J. Barát, Gy. Muszély - Theoretical and Experimental Investigation of a Structure - Adaptive Pattern Recognition Algorithm.....	79

Wydawnictwa Czasopism Technicznych NOT - Polska

Zakład Poligraficzny WCT NOT. Zam. 57/69.

ОБУЧАЮЩИЕСЯ СИСТЕМЫ УПРАВЛЕНИЯ

Я.З. Цыпкин, Г.К. Кельманс, Л.Е. Эпштейн

Институт автоматики и телемеханики /технической кибернетики/
Москва /СССР/

В в е д е н и е

Обучающиеся автоматические системы характеризуются способностью улучшать свое функционированное поведение, свойства в ходе работы. Настоящая статья посвящена обсуждению принципа построения обучающихся автоматических систем, выяснению их возможностей и особенностей.

Необходимость в использовании обучающихся автоматических систем возникает в тех случаях, когда характеристики наблюдаемых ситуаций /соответствующих сигналам, подлежащим опознанию, либо состояниям управляемых объектов/ заранее неизвестны, т.е. когда априорная информация о наблюдаемых ситуациях мала.

Принцип работы обучающихся автоматических систем основан на алгоритмах обучения классификации наблюдаемых ситуаций /образов/. Поэтому в работе большое внимание уделяется методу получения этих алгоритмов на основе минимизации функционала типа среднего риска ошибочной классификации. Этот метод вытекает из общей теории адаптации и обучения^[1].

Показано, что при частных видах функций потерь могут быть получены как известные к настоящему времени, так и новые алгоритмы обучения и самообучения.

Полученные алгоритмы применены для построения обучающихся систем управления и обучающегося приемника импульсных сигналов.

Приводятся результаты экспериментального исследования самообучающегося приемника.

1. Виды обучения

В основу построения обучающихся автоматических систем нами положены алгоритмы, классификации наблюдаемых ситуаций. Разбиение ситуаций на классы осуществляется на основе некоторого решающего правила. Методы определения этого решающего правила

существенно зависят от объема априорной информации о классифицируемых ситуациях.

При достаточной априорной информации, т.е. когда нам известны вероятностные характеристики ситуаций, решающее правило определяется классическими методами теории статистических решений [2,3]. В этом случае обучение отсутствует. Необходимость в обучении возникает при неполной, недостаточной априорной информации, т.е. когда вероятностные характеристики ситуаций заранее неизвестны. Если имеется дополнительная информация, состоящая в том, что заранее известна принадлежность ситуаций из некой последовательности к тому или иному классу, то решающее правило находится на основе этой "тренировочной" последовательности с помощью алгоритмов обучения с поощрением [4,6].

Если же и такая дополнительная информация отсутствует, то решающее правило должно быть найдено только на основе наблюдаемых ситуаций с помощью алгоритмов, которые уместно называть алгоритмами самообучения [7,8]. Таким образом, мы сталкиваемся с двумя видами обучения: обучения с поощрением и обучением без поощрения, или самообучением.

Решающие правила для всех перечисленных случаев ранее получались в результате рассмотрения самостоятельных и, на первый взгляд, не связанных между собой задач. Тем не менее все известные, а также новые решающие правила могут быть получены регулярным путем из условий минимума общего критерия качества классификации типа среднего риска.

К формулировке критерия оптимальности классификации ситуаций и получению условий его минимума мы сейчас и перейдем.

2. Критерий классификации

Для построения решающего правила необходимо прежде всего сформулировать критерий качества классификации, экстремум которого и должен определить это правило.

Предположим, что ситуации x из пространства X появляются случайным образом и каждая возникающая ситуация относится к одному из N неизвестных нам классов $X_1^0, X_2^0, \dots, X_N^0$. Причем, при разных появлениях одна и та же ситуация x может быть отнесена к разным классам.

Разобьем пространство ситуаций X на N областей и будем относить ситуации $x \in X_i$, возникающую в любой момент времени, к классу X_i^0 . Задача классификации состоит в оптимальном в некотором смысле разбиении пространства ситуаций X на области $X_1, X_2, \dots, X_N$.

Введем функции потерь $F_{km}(x, u_1, u_2, \dots, u_N)$ ($k, m = 1, 2, \dots, N$), где $u_1, u_2, \dots, u_N$ — некоторые параметры. Каждая функция $F_{km}(x, u_1, u_2, \dots, u_N)$ оценивает потери, возникающие при отнесении ситуации x из класса X_k^0 к классу X_m^0 или, иначе, при попадании ситуации из класса X_k^0 в область X_m .

Пусть также

P_k — вероятность наступления ситуаций класса X_k^0 ; $\rho(x/k) = \rho_k(x)$ — условная плотность вероятности ситуаций класса X_k^0 . Тогда средний риск неправильной классификации можно представить в следующем виде:

$$R = \sum_{k=1}^N \sum_{m=1}^N \int_{X_m} F_{km}(x, u_1, u_2, \dots, u_N) P_k \rho_k(x) dx \quad /2.1/$$

Наилучшая классификация соответствует такому выбору параметров $u_1, u_2, \dots, u_N$ и областей $X_1, X_2, \dots, X_N$, при котором средний риск /2.1/ минимален.

3. Условия минимума среднего риска

Средний риск представляет собой функционал от границ $\Lambda_{km}(x)$ между соседними областями X_k и X_m и набора векторов $u_1, u_2, \dots, u_N$. Для нахождения условия минимума среднего риска мы, как и в [8], используем классические методы вариационного исчисления [9], применив их теперь к более сложному функционалу /2.1/. Эти условия минимума среднего риска могут быть представлены в такой форме:

$$\sum_{k=1}^N \sum_{m=1}^N \int_{X_m} \nabla_{u_\ell} F_{km}(x, u_1, u_2, \dots, u_N) P_k \rho_k(x) dx = 0, (\ell = 1, 2, \dots, N) \quad /3.1/$$

и для всех x , принадлежащих границе Λ_{sm} между соседними областями X_s и X_m .

$$f_{sm}(x, u_1, u_2, \dots, u_N) = \sum_{k=1}^N [F_{ks}(x, u_1, u_2, \dots, u_N) - F_{km}(x, u_1, u_2, \dots, u_N)] P_k \rho_k(x) = 0 \quad /3.2/$$

Уравнение /3.2/ представляет собой поверхность, разделяющую соседние области X_s и X_m . Поэтому целесообразно функцию

$f_{sm}(\cdot)$ назвать разделяющей. Знаки разделяющих функций при $x \in \Lambda_{sm}$ для всех $s, m=1, 2, \dots, N$ позволяют различать области.

При определенных предположениях [7] решающее правило можно сформулировать следующим образом

$$x \in X_m \quad (m=1, 2, \dots, N), \text{ если } f_{ms}(x, u_1, u_2, \dots, u_N) < 0 \quad /3.3/$$

для всех $s \neq m$ ($s=1, 2, \dots, N$). В частном случае для двух классов средний риск R , условия его минимума и решающие правила принимают вид:

$$R = \int_{X_1} F_{11}(x, u_1, u_2) P_1 p_1(x) dx + \int_{X_1} F_{21}(x, u_1, u_2) P_2 p_2(x) dx + \\ + \int_{X_2} F_{12}(x, u_1, u_2) P_1 p_1(x) dx + \int_{X_2} F_{22}(x, u_1, u_2) P_2 p_2(x) dx \quad /3.4/$$

$$\nabla_{u_i} R = \int_{X_1} \nabla_{u_i} F_{11}(x, u_1, u_2) P_1 p_1(x) dx + \int_{X_1} \nabla_{u_i} F_{21}(x, u_1, u_2) P_2 p_2(x) dx + \\ + \int_{X_2} \nabla_{u_i} F_{12}(x, u_1, u_2) P_1 p_1(x) dx + \int_{X_2} \nabla_{u_i} F_{22}(x, u_1, u_2) P_2 p_2(x) dx = 0 \quad /3.5/$$

$$x \text{ относится к классу } X_1^0, \text{ если } f_{12}(x, u_1, u_2) < 0 \quad /3.6/$$

$$x \text{ относится к классу } X_2^0, \text{ если } f_{12}(x, u_1, u_2) > 0$$

Дальнейшая задача состоит в окончательном определении из этих условий решающих правил.

4.7 Формулировки задачи

В зависимости от характера априорной и текущей дополнительной информации возможны различные формулировки задач, связанных с определением решающего правила, минимизирующего средний риск /2.1/.

Задача I. Вероятности P_k появления ситуации классов X_k^0 и условные ~~яне~~ плотности вероятности ситуаций $p_k(x)$ в этих классах известны ($k=1, 2, \dots, N$). Известны также функции потерь $F_{km}(x, u_1, \dots, u_N) = F_{km}(x)$. В этом случае нет необходимости в использовании обучения для получения решающего правила.

Задача II. Вероятности P_k и условные плотности вероятности ситуаций $p_k(x)$ известны ($k=1, 2, \dots, N$). Функции же потерь заданы с точностью до параметров $u_1, u_2, \dots, u_N$. В этом случае уже возникает необходимость использовать для оценки параметров $u_1, u_2, \dots, u_N$, определяющих решающие правила, обучение на основе дополнительной текущей информации о

принадлежности возникающих ситуаций к тем или иным классам, т.е. на основе поощрений и наказаний.

Задача Па. Вероятности P_k и плотности вероятности $\rho_k(x)$ ($k=1,2,\dots,N$) неизвестны. Функции потерь $F_{km}(x, u_1, u_2, \dots, u_N)$ заданы с точностью до параметров $u_1, u_2, \dots, u_N$. В этом случае для оценки оптимальных параметров $u_1, u_2, \dots, u_N$ необходимо восстанавливать некоторые статистические характеристики ситуаций. Для этого нужно, как и ранее, использовать дополнительную текущую информацию о принадлежности возникающих ситуаций к соответствующим классам.

В перечисленных выше задачах предполагалось, что существует некоторая классификация ситуаций. И задача состояла в получении приближения к этой классификации. В случае отсутствия дополнительной текущей информации о ситуациях такая постановка задачи, естественно, теряет смысл.

С этим связаны изменения физической сути соответствующих составляющих среднего риска /2.1/ в задаче Ш.

Задача Ш. Вероятности P_k появления ситуаций из областей X_k /а не из классов X_k^0 , как было ранее/, и условные плотности вероятности ситуаций $\rho_k(x)$ в этих областях, неизвестны ($k=1,2,\dots,N$).

Функции потерь $F_{km}(x, u_1, u_2, \dots, u_N) = F_m(x, u_1, u_2, \dots, u_N)$ /для всех $k=1,2,\dots,N$ /, выражающие теперь штраф за отнесение ситуаций x к классу X_m ($m=1,2,\dots,N$), заданы с точностью до параметров $u_1, u_2, \dots, u_N$. При этих условиях оценка решающего правила должна быть основана только на обработке наблюдаемых ситуаций без использования текущей информации о принадлежности ситуаций к соответствующим классам.

Таким образом, в зависимости от степени полноты априорной информации и наличия дополнительной текущей информации о наблюдаемых ситуациях следует различать распознавание или классификацию без обучения /задача I/; с обучением при использовании дополнительной текущей информации - обучением с поощрением /задачи П, Па/; с обучением при отсутствии дополнительной информации - обучением без поощрения, или самообучением /задача Ш/.

Перейдем теперь к определению решающих правил. Для простоты и обозримости результатов мы в основном будем рассматривать случай двух классов.

5. Определение решающего правила без обучения /задача I/

Положим

$$\begin{aligned} F_{11}(x, u_1, u_2) &= F_{22}(x, u_1, u_2) = 0 \\ F_{12}(x, u_1, u_2) &= F_1(x) \\ F_{21}(x, u_1, u_2) &= F_2(x) \end{aligned} \quad /5.1/$$

Это означает, что при правильной классификации ситуации x потери равны нулю, а потери при ошибках 1-го и 2-го рода зависят от ситуации. Тогда условие минимума среднего риска /3.4/ принимает вид:

$$f_{12}(x) = P_2 p_2(x) F_2(x) - P_1 p_1(x) F_1(x) = 0 \quad /5.2/$$

для всех $x \in \Lambda_{12}$. $f_{12}(x)$ является разделяющей функцией, поэтому решающее правило принимает вид:

$$\begin{aligned} x \in X_1 & \quad , \text{ если } f_{12}(x) < 0 \\ x \in X_2 & \quad , \text{ если } f_{12}(x) > 0 \end{aligned} \quad /5.3/$$

Решающее правило можно сформулировать через коэффициент правдоподобия $\frac{p_1(x)}{p_2(x)}$, а именно:

$$\begin{aligned} x \in X_1 & \quad , \text{ если } \frac{p_1(x)}{p_2(x)} > h(x) \\ x \in X_2 & \quad , \text{ если } \frac{p_1(x)}{p_2(x)} < h(x) \end{aligned} \quad /5.4/$$

где $h(x) = \frac{P_2 F_2(x)}{P_1 F_1(x)}$ представляет собой порог, который в данном случае зависит от x . В частном случае, когда

$$F_{12}(x) = w_{12} = \text{const.} \quad , \quad F_{21}(x) = w_{21} = \text{const.}$$

$$h(x) = h = \frac{P_2 \cdot w_{21}}{P_1 \cdot w_{12}} = \text{const.} \quad /5.5/$$

Именно этот случай обычно и рассматривается в статистической теории обнаружения [2, 3]. В любом ином случае при неполной информации решающее правило может быть определено лишь с помощью алгоритмов обучения.

6. Алгоритмы обучения с поощрением

Для получения алгоритмов обучения с поощрением преобразуем условия минимума /3.1/, /3.2/ среднего риска R к более

удобной форме:

$$M\Phi_e(x, u_1, u_2) = 0, \quad /6.1/$$

$$\begin{aligned} x \in X_1 & \quad , \text{ если } f_{12}(x, u_1, u_2) < 0 \\ x \in X_2 & \quad , \text{ если } f_{12}(x, u_1, u_2) > 0 \end{aligned} \quad /6.2/$$

где

$$\begin{aligned} \Phi_e(x, u_1, u_2) &= \nabla_{u_e} F_{km}(x, u_1, u_2) \quad , \text{ если ситуация клас-} \\ & \quad (\kappa, \ell, m = 1, 2) \quad \text{са } X_{\kappa}^0 \text{ относится} \\ & \quad \text{к классу } X_m^0 \end{aligned} \quad /6.3/$$

Теперь для определения решающего правила можно воспользоваться итеративными алгоритмами обучения [I].

Вид алгоритмов обучения зависит от полноты априорной информации.

Задача II. Предположим, что вероятностные характеристики ситуаций известны. Неизвестными являются параметры u_1, u_2 функций потерь $F_{ik}(x, u_1, u_2)$ ($i \neq \kappa; i, \kappa = 1, 2$). Как и ранее, будем полагать $F_{mm}(x, u_1, u_2)$ ($m = 1, 2$) равными нулю. Тогда разделяющая функция $f_{12}(x)$ принимает вид:

$$f_{12}(x) = P_2 p_2(x) \cdot F_{21}(x, u_1, u_2) - P_1 p_1(x) \cdot F_{12}(x, u_1, u_2) \quad /6.4/$$

Применяя вероятностные итеративные алгоритмы к /6.1/, получим:

$$u_e[n] = u_e[n-1] - \delta_e[n] \cdot \nabla_{u_e} F_{12}(x[n], u_1[n-1], u_2[n-1]),$$

если $x[n]$ - ситуация класса X_1^0 и

$$F_{21}(x[n], u_1[n-1], u_2[n-1]) \cdot P_2 p_2(x[n]) - F_{12}(x[n], u_1[n-1], u_2[n-1]) \cdot P_1 p_1(x[n]) > 0 \quad /6.5/$$

$$u_e[n] = u_e[n-1] - \delta_e[n] \cdot \nabla_{u_e} F_{21}(x[n], u_1[n-1], u_2[n-1]),$$

если $x[n]$ - ситуация класса X_2^0 и

$$F_{21}(x[n], u_1[n-1], u_2[n-1]) \cdot P_2 p_2(x[n]) - F_{12}(x[n], u_1[n-1], u_2[n-1]) \cdot P_1 p_1(x[n]) < 0$$

$$u_e[n] = u_e[n-1],$$

- в остальных случаях ($\ell = 1, 2$).

Задача Па. Пусть теперь вероятностные характеристики ситуаций неизвестны. В этом случае также можно было бы воспользоваться алгоритмами обучения, приведенными выше, если бы в процессе обучения удавалось восстанавливать вероятностные характеристики ситуаций, входящие в выражение для разделяющей функции

$$f_{12}(x, u_1, u_2) = F_{21}(x, u_1, u_2) \cdot P_2 \rho_2(x) - F_{12}(x, u_1, u_2) P_1 \rho_1(x) \quad /6.6/$$

Это можно сделать с помощью известных алгоритмов восстановления вероятностей и плотностей распределения случайных величин [9, 10]. Однако, в этом нет необходимости. Мы остановимся здесь на ином, более простом способе, основанном на использовании, по существу, этих же алгоритмов, но для восстановления не составляющих разделяющей функции, а непосредственно самой разделяющей функции. Зафиксируем параметры u_1, u_2 , тогда

$$f_{12}(x, u_1, u_2) = f_{12}(x) \quad /6.7/$$

Будем аппроксимировать функцию $f_{12}(x)$ отрезком ортонормированного ряда

$$\hat{f}(x) = c^T \varphi(x) \quad /6.8/$$

так, чтобы функционал

$$I(c) = \int_X [f_{12}(x) - c^T \varphi(x)]^2 dx \quad /6.9/$$

был минимален. В /6.8/ и /6.9/ знак T означает транспонирование вектора c . Дифференцируя функцию $I(c)$ по c и используя свойство ортонормированности компонент вектор-функции $\varphi(x)$, находим оптимальные значения c :

$$c = M \Phi(x), \quad /6.10/$$

где

$$\Phi(x) = \begin{cases} \varphi(x) \cdot F_{12}(x) & , \text{ если } x - \text{ ситуация класса } X_1^0 \\ -\varphi(x) \cdot F_{21}(x) & ; \text{ если } x - \text{ ситуация класса } X_2^0 \end{cases} \quad /6.11/$$

Рассматривая условия /6.10/ как уравнение регрессии, получаем на основании вероятностных итеративных методов I, II

алгоритм:

$$c[n] = c[n-1] - \delta[n] \cdot \{c[n-1] - \varphi(x[n]) \cdot F_{12}(x[n])\}, \quad /6.12/$$

если $x[n]$ - ситуация класса X_1^0 ,

$$c[n] = c[n-1] - \delta[n] \cdot \{c[n-1] + \varphi(x[n]) \cdot F_{21}(x[n])\},$$

если $x[n]$ - ситуация класса X_2^0 . Применяя для решения задачи ПА /6.10/ при произвольных параметрах u_1, u_2 , а также используя /6.1/, получим необходимые условия относительно параметров u_1, u_2, c .

$$M \Phi_\ell(x, u_1, u_2) = 0, \quad (\ell = 1, 2)$$

/6.13/

$$M \Phi(x, u_1, u_2) = c$$

$c^T \cdot \varphi(x) = 0$ для всех $x \in \Lambda_{12}$, где функции $\Phi_\ell(x, u_1, u_2)$ и $\Phi(x, u_1, u_2)$ определяются выражениями соответственно /6.3/ и /6.11/.

Из системы уравнений /6.13/ вытекают следующие алгоритмы обучения

$$u_\ell[n] = u_\ell[n-1] - \delta_\ell[n] \cdot \nabla_{u_\ell} F_{12}(x[n], u_1[n-1], u_2[n-1])$$

$$c[n] = c[n-1] - \delta[n] \cdot \{c[n-1] - \varphi(x[n]) \cdot F_{12}(x[n], u_1[n-1], u_2[n-1])\},$$

если $x[n]$ - ситуация класса X_1^0 , и $c^T[n-1] \cdot \varphi(x[n]) > 0$

$$u_\ell[n] = u_\ell[n-1] - \delta_\ell[n] \cdot \nabla_{u_\ell} F_{21}(x[n], u_1[n-1], u_2[n-1])$$

$$c[n] = c[n-1] - \delta[n] \cdot \{c[n-1] + \varphi(x[n]) \cdot F_{21}(x[n], u_1[n-1], u_2[n-1])\} \quad /6.14/$$

если $x[n]$ - ситуация класса X_2^0 , и $c^T[n-1] \cdot \varphi(x[n]) < 0$

$$c[n] = c[n-1] - \delta[n] \cdot \{c[n-1] + (-1)^K \cdot \varphi(x[n]) \cdot F_{km}(x[n], u_1[n-1], u_2[n-1])\}$$

если $x[n]$ - ситуация класса X_k^0 ($k \neq m$; $k, m = 1, 2$)

$u_\ell[n] = u_\ell[n-1]$, ($\ell = 1, 2$) - в остальных случаях. В наиболее распространенном случае, лежащем в основе статистической теории обнаружения [1, 2],

$$F_{12}(x, u_1, u_2) = w_{12}; \quad F_{21}(x, u_1, u_2) = w_{21} \quad /6.15/$$

При этом алгоритм оценки коэффициентов c разделяющей функции $\hat{f}(x) = c^T \cdot \varphi(x)$, минимизирующей средний риск

$$R = \int_{X_1} w_{21} P_2 p_2(x) dx + \int_{X_2} w_{12} P_1 p_1(x) dx \quad /6.16/$$

будет иметь следующий вид:

$$\begin{aligned}
 &C[n] = C[n-1] - \delta[n] \cdot \{C[n-1] - \varphi(x[n]) \cdot w_{12}\}, \\
 &\text{если } x[n] - \text{ситуация класса } X_1^0; \quad /6.17/ \\
 &C[n] = C[n-1] - \delta[n] \cdot \{C[n-1] + \varphi(x[n]) \cdot w_{21}\}, \\
 &\text{если } x[n] - \text{ситуация класса } X_2^0.
 \end{aligned}$$

7. Алгоритмы обучения без поощрения /алгоритмы самообучения/

В этом случае нам неизвестны ни вероятностные характеристики ситуаций x , ни дополнительная информация о принадлежности ситуаций к классам. Иначе говоря, используются только наблюдения ситуаций, принадлежность которых к тем или иным классам неизвестна. Принимая во внимание обозначение § 4, запишем средний риск R при $N=2$ так:

$$R = \sum_{k=1}^2 \int_{X_k} F_k(x, u_1, u_2) p_k p_k(x) dx \quad /7.1/$$

Условие минимума получается из /6.1/, /6.2/, /6.3/ заменой условия принадлежности ситуаций к реальным классам X_k^0 на условие ($x \in X_k$) принадлежности ситуации x к области X_k :

$$M \Phi_e(x, u_1, u_2) = 0, \quad /7.2/$$

где

$$\Phi_e(x, u_1, u_2) = \nabla_{u_e} F_k(x, u_1, u_2), \quad \text{если } x \in X_k \quad /7.3/$$

($k, e \neq 1, 2$)

и

$$\begin{aligned}
 &x \in X_1, \quad \text{если } F_1(x, u_1, u_2) - F_2(x, u_1, u_2) < 0 \\
 &x \in X_2, \quad \text{если } F_1(x, u_1, u_2) - F_2(x, u_1, u_2) > 0 \quad /7.4/
 \end{aligned}$$

Для определения алгоритмов обучения, определяющих параметры u_1, u_2 , применим к /7.2/ вероятностные итеративные алгоритмы. Тогда получим:

$$u_e[n] = u_e[n-1] - \delta[n] \cdot \nabla_{u_e} F_1(x[n], u_1[n-1], u_2[n-1]),$$

если

$$\begin{aligned}
 &F_1(x[n], u_1[n-1], u_2[n-1]) - F_2(x[n], u_1[n-1], u_2[n-1]) < 0 \quad /7.5/ \\
 &u_e[n] = u_e[n-1] - \delta[n] \cdot \nabla_{u_e} F_2(x[n], u_1[n-1], u_2[n-1]),
 \end{aligned}$$

если $F_1(x[n], u, [n-1], u_2[n-1]) - F_2(x[n], u, [n-1], u_2[n-1]) > 0$

Эти алгоритмы, естественно, совпадают с алгоритмами, полученными в [7].

8. Частные случаи алгоритмов

Полученные выше алгоритмы, хотя по внешнему виду и отличаются от ранее известных алгоритмов классификации ситуаций, тем не менее содержат последние как частный случай. Покажем это для задачи обучения с поощрением.

Будем искать разделяющую функцию в форме

$$\hat{y} = u^T \varphi(x) \quad /8.1/$$

где u - вектор неизвестных параметров,

$\varphi(x)$ - вектор-функция, компоненты которой линейно независимы. Выберем в основном критерий /7.1/ функции потерь следующим образом:

$$\begin{aligned} F_{12}(x, u) &= F_{11}(x, u) = F(1 - u^T \varphi(x)) \\ F_{21}(x, u) &= F_{22}(x, u) = F(-1 - u^T \varphi(x)) \end{aligned} \quad /8.2/$$

где $F(\cdot)$ - некоторая выпуклая функция. Тогда

$$\begin{aligned} R = \int_{x_1 = \{x: u^T \varphi(x) < 0\}} F(1 - u^T \varphi(x)) p_1(x) dx + \int_{x_2 = \{x: u^T \varphi(x) > 0\}} F(-1 - u^T \varphi(x)) p_2(x) dx + \\ + \int_{x_1} F(-1 - u^T \varphi(x)) p_2(x) dx + \int_{x_2} F(1 - u^T \varphi(x)) p_1(x) dx \end{aligned} \quad /8.3/$$

Вводя смешанную плотность распределения ситуаций

$$\rho(x) = p_1 p_1(x) + p_2 p_2(x) \quad /8.4/$$

можно представить средний риск /8.3/ в виде

$$R = \int_x F(y - u^T \varphi(x)) \rho(x) dx = MF(y - u^T \varphi(x)) \quad /8.5/$$

где

$$y = \begin{cases} 1 & , \text{ если } x \text{ относится к классу } X_1^0 \\ -1 & , \text{ если } x \text{ относится к классу } X_2^0 \end{cases} \quad /8.6/$$

- указания учителя /дополнительная информация/ о принадлежности возникшей ситуации к тому или иному классу. Именно в такой форме и рассматривались функционалы в задачах обучения распознаванию [4-6].

Алгоритм обучения в этом случае имеет вид:

$$\begin{aligned} u[n] &= u[n-1] - \delta[n] \cdot \nabla_u F(1 - u^T[n-1] \cdot \varphi(x[n])), \\ \text{если } x[n] &\text{ - ситуация класса } X_1^0; \\ u[n] &= u[n-1] - \delta[n] \cdot \nabla_u F(-1 - u^T[n-1] \cdot \varphi(x[n])), \end{aligned} \quad /8.7/$$

если $x[n]$ - ситуация класса X_2° . Алгоритмы /8.7/ можно записать и в более компактной форме, используя /8.6/,

$$u[n] = u[n-1] - \gamma[n] \cdot \nabla_u F(y[n] - u^T[n-1] \cdot \varphi(x[n]))$$

Такие алгоритмы в работе [5] и были получены непосредственно из функционала /8.5/.

9. Об оптимальных алгоритмах

Алгоритмы обучения, приведенные выше, сходятся при выполнении определенных условий, налагаемых на величины $\gamma[n]$ и функции потерь [1]. В тех случаях, когда число наблюдаемых ситуаций ограничено, целесообразно подчинить выбор $\gamma[n]$ условию наилучшего в некотором смысле использования этой информации, т.е. строить оптимальные алгоритмы [12, 13]. Здесь есть 2 возможности [13]:

одна из них основана на многократном использовании неоптимальных алгоритмов с относительно простым $\gamma[n]$ типа $\gamma[n] = \frac{\alpha}{n}$ при периодическом использовании известной последовательности ситуаций;

другая - на специальном выборе $\gamma[n]$, минимизирующем некоторый эмпирический функционал. Так, в частности, для линейных алгоритмов типа /6.12/, /6.17/ легко показать, например, как это было сделано в [13, 14], что оптимальное значение

$$\gamma_{\text{опт}}[n] = \frac{1}{n + \frac{\sigma^2}{v^2[0]}} ; \quad /9.1/$$

где σ^2 - дисперсия x , $v^2[0]$ - начальная дисперсия оценки. Если априорная информация относительно начального значения дисперсии оценки отсутствует, то $v^2[0] = \infty$ и получаем

$$\gamma_{\text{опт}}[n] = \frac{1}{n} . \quad /9.2/$$

Что касается общих случаев, то определение оптимальных значений весьма громоздко, и мы здесь этих вопросов касаться не будем.

10. Обучающиеся системы управления

Теория оптимальных систем управления дает возможность получить закон изменения управляющего воздействия $u_y(t)$ как функцию времени. Что же касается синтеза оптимального закона управления, т.е. получения зависимости u_y от фазовых

координат, то эта задача далека от приемлемого решения в сколь-нибудь сложных случаях [3, 15]. Для оптимальных по быстродействию систем

$$u_y = 1, \quad \text{либо} \quad u_y = -1 \quad /10.1/$$

Поэтому задачу синтеза оптимальной по быстродействию системы можно рассматривать как задачу классификации вектора фазовых координат $x = (x_1, x_2, \dots, x_n)$ на две категории /10.1/. Представим неизвестное нам уравнение разделяющей поверхности в виде

$$\hat{f}(x) = u^T \varphi(x) = 0 \quad /10.2/$$

Тогда

$$u_y = \begin{cases} 1 & , \text{если } \hat{f}(x) \geq 0 \\ -1 & , \text{если } \hat{f}(x) < 0 \end{cases} \quad /10.3/$$

По результатам определения оптимальных управляющих воздействий $u_{\text{опт.}}(t)$ и траектории фазовых координат $x_{\text{опт.}}(t)$ при заданных начальных условиях $x(0)$ можно составить тренировочную последовательность, т.е. соответствие между $x_{\text{опт.}}$ и $u_{\text{опт.}}$. Классификация /10.3/ осуществляется с помощью алгоритмов обучения, приведенных в § 6. Таким образом, в таких обучающихся системах управления на основе тренировочной последовательности восстанавливается так называемая поверхность переключения. Реализация подобных систем на основе адаптивных линейных элементов рассматривалась в работах [16 - 18]. Допустим, что мы имеем уникальный оптимальный по быстродействию регулятор, который в зависимости от координат объекта вырабатывал бы оптимальное управление /10.1/. Этот оптимальный регулятор можно использовать для настройки параметров "обычных" типовых регуляторов, входящих в состав простых обучающихся систем управления. В процессе обучения настраиваемые параметры изменяются так, чтобы с течением времени "обычный" регулятор наилучшим образом /в смысле минимума ошибочных управляющих воздействий/ выполняли бы функции дорогостоящего "оптимального" регулятора [1, 16].

Иные возможности построения обучающихся систем управления основаны на дискретизации пространства входных сигналов системы и определении или исследовании уже известных оптимальных управлений для каждой из получаемых областей [19, 20].



II. Самообучающиеся приемники

Задача самообучающегося приемника состоит в классификации принимаемых сигналов. В тех случаях, когда сигналы представляются скалярной функцией времени, мы можем воспользоваться для построения такого приемника алгоритмами самообучения распознаванию ситуаций, внося при этом те упрощения, которые обуславливаются рассмотрением одномерного случая. Так, при квадратичных функциях потерь критерий качества запишется в виде:

$$R = \int_{x_1} (u_1 - x)^2 p(x) dx + \int_{x_2} (u_2 - x)^2 p(x) dx \quad /II.1/$$

В этой задаче целесообразнее использовать не дискретные, а непрерывные алгоритмы, которые могут быть получены из приведенных выше дискретных путем соответствующего предельного перехода /см. также [8] /. В данном случае эти алгоритмы самообучения могут быть представлены в такой форме:

$$\begin{aligned} \frac{du_1(t)}{dt} &= -\gamma_1(t) \cdot [x(t) - u_1(t)] \cdot f(x(t) - x^0) \\ \frac{du_2(t)}{dt} &= -\gamma_2(t) \cdot [x(t) - u_2(t)] \cdot f(x^0 - x(t)), \end{aligned} \quad /II.2/$$

где x^0 - порог, равный

$$x^0 = \frac{u_1 + u_2}{2} \quad /II.3/$$

и

$$f(z) = \begin{cases} 1 & , \text{если } z \geq 0 \\ 0 & , \text{если } z < 0 \end{cases} \quad /II.4/$$

Оптимальные значения $\gamma_1(t)$ и $\gamma_2(t)$ можно представить в такой форме:

$$\gamma_1(t) = \frac{1}{\int_0^t x^2(\tau) \cdot f(x(\tau) - x^0(t)) d\tau} \quad /II.5/$$

$$\gamma_2(t) = \frac{1}{\int_0^t x^2(\tau) \cdot f(x^0(t) - x(\tau)) d\tau} \quad /II.6/$$

где $t > \varepsilon > 0$.

Структурная схема приемника представлена на рис. 1. Она осуществлена на основе элементов ЭМУ-8 и специально разработанных блоков /полупроводниковые бесконтактные реле, γ - элемент и управляемый потенциометр/. Ниже приводятся результаты экспериментального исследования самообучающегося приемника для осуществления классификации входных сигналов. В качестве таковых использовались импульсные сигналы, амплитуда которых менялась по различным законам, и импульсные сигналы на фоне помех. На осциллограмме рис. 2 изображены процессы обучения без дополнительной информации извне, где обозначено:

- 1 - входной сигнал приемника,
- 2 - выход сигналов, относящихся к 1-ому классу x_1 ,
- 3 - выход сигналов, относящихся ко 2-ому классу x_2 ,
- 4 - процесс установления порога. 0

Время обучения можно судить по длительности установления порога. Для рассматриваемых случаев процесс обучения длится 3 - 7 периодов. Отношение сигнал - помеха /на осциллограммах рис. 2, а-с/ имеют порядок 0,85-0,4. Естественно, что качество работы такого принципа единственно зависит от того, насколько четко разделяются классифицируемые сигналы. Подобная схема обучающегося приемника была использована для установки параметра срабатывания /порога/ в макете линии телетайпной связи с помехами и обеспечивают минимальное число ошибок в принимаемом самообучении при заранее неизвестной статистике помех. Подобные приемники могут найти применение и в ряде других областей /телеуправление, контроль и т.п./.

З а к л ю ч е н и е

В работе развита теория обучающихся автоматических систем. Из условий минимума среднего риска ошибочной классификации ситуаций с помощью вероятностных итеративных алгоритмов получены алгоритмы классификации ситуаций как при полной, так и при недостаточной информации о классифицируемых ситуациях. В последнем случае мы приходим к алгоритмам обучения с поощрением или без поощрения. Из этих алгоритмов в качестве частных случаев следуют известные ранее алгоритмы обучения и самообучения. Приводятся примеры применения алгоритмов обучения для построения обучающихся систем управления и самообучающихся приемников.

Л и т е р а т у р а

1. Цыпкин Я.З., Адаптация и обучение в автоматических системах. "Наука", М. 1968.
2. Миддлтон Л., Введение в статистическую теорию связи, т. I, II. Изд-во "Сов.радио", 1962.
3. Фельдбаум А.А., Основы теории оптимальных автоматических систем. "Наука", 1966.
4. Якубович В.А., Некоторые общие теоретические принципы построения обучаемых опознающих систем I. Вычислительная

x_1 / При воздействии сигналов меняющейся амплитуды /рис.2/ выходные сигналы взяты нормированными.

- техника и вопросы программирования, вып. 4, издание ЛГУ, 1965.
5. Айзерман М.А., Браверман Э.М., Розоноэр Л.И., Вероятностная задача об обучении автоматов распознаванию классов и метод потенциальных функций. Автоматика и телемеханика, 1966, XXV, № 1.
 6. Девятериков И.П., Пропой А.И., Цыпкин Я.З., О рекуррентных алгоритмах обучения распознаванию образов. Автоматика и телемеханика, 1967, XXV, № 1.
 7. Браверман Э.М., Метод потенциальных функций в задаче обучения машины распознаванию образов без учителя. Автоматика и телемеханика, 1966, XXV, № 10.
 8. Цыпкин Я.З., Кельманс Г.К., Рекуррентные алгоритмы самообучения. Изв. АН СССР, Техническая кибернетика, 1967, №5.
 9. Гельфанд И.М., Фомин С.В., Вариационное исчисление, ФМ, 1961.
 10. Цыпкин Я.З., Применение метода стохастической аппроксимации к оценке неизвестной плотности распределения по наблюдениям. Автоматика и телемеханика, 1966, XXV, № 3.
 11. Цыпкин Я.З., Об алгоритмах оценки плотности распределения и их моментов по наблюдениям. Автоматика и телемеханика, 1967, № 7.
 12. Стратонович Р.Л., Существует ли теория синтеза оптимальных адаптивных самообучающихся и самонастраивающихся систем? Автоматика и телемеханика, 1968, № 1.
 13. ^{Цыпкин Я.З.} А все же существует ли теория синтеза оптимальных адаптивных систем? Автоматика и телемеханика, 1968, № 1.
 14. Nicolic Z.T., Fu K.S., A mathematical model of learning in an unknown random environment. Proc. of the National Electronics Conference, vol. XXII, 1966.
 15. Понтрягин Л.С., Болтянский В.Г., Гамкредидзе Р.В., Мищенко Е.Ф., Математическая теория оптимальных процессов. ФМ, 1961.
 16. Widrow B., Smith F.W., Pattern recognizing control systems. Computer and Information Sciences, Symposium Proc. Spartan, 1964.
 17. Smith F.W., A trainable nonlinear function generator. IEEE Trans. on Automatic Control, 1966, vol. AC-11, No. 2.
 18. Mendel J.M., Zaplac I.T., Realization a suboptimal controller by off-line training technique. 1967 Joint Automatic Control Conference.

19. Тейлор В.К., Самонастраивающееся устройство управления, использующее распознавание образов. Труды II Международного конгресса ИФАК. "Наука", 1965.
20. Чеховой Ю.Н., Керкеснер И.П., Обучающаяся система автоматического управления, использующая распознавание ситуаций. Труды III Всесоюзного совещания по автоматическому управлению /технической кибернетики/, "Наука", 1967.

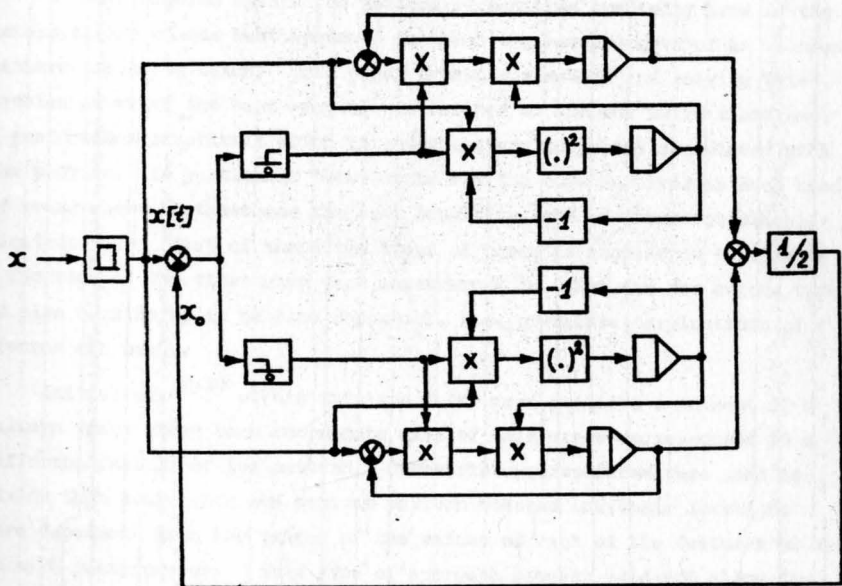
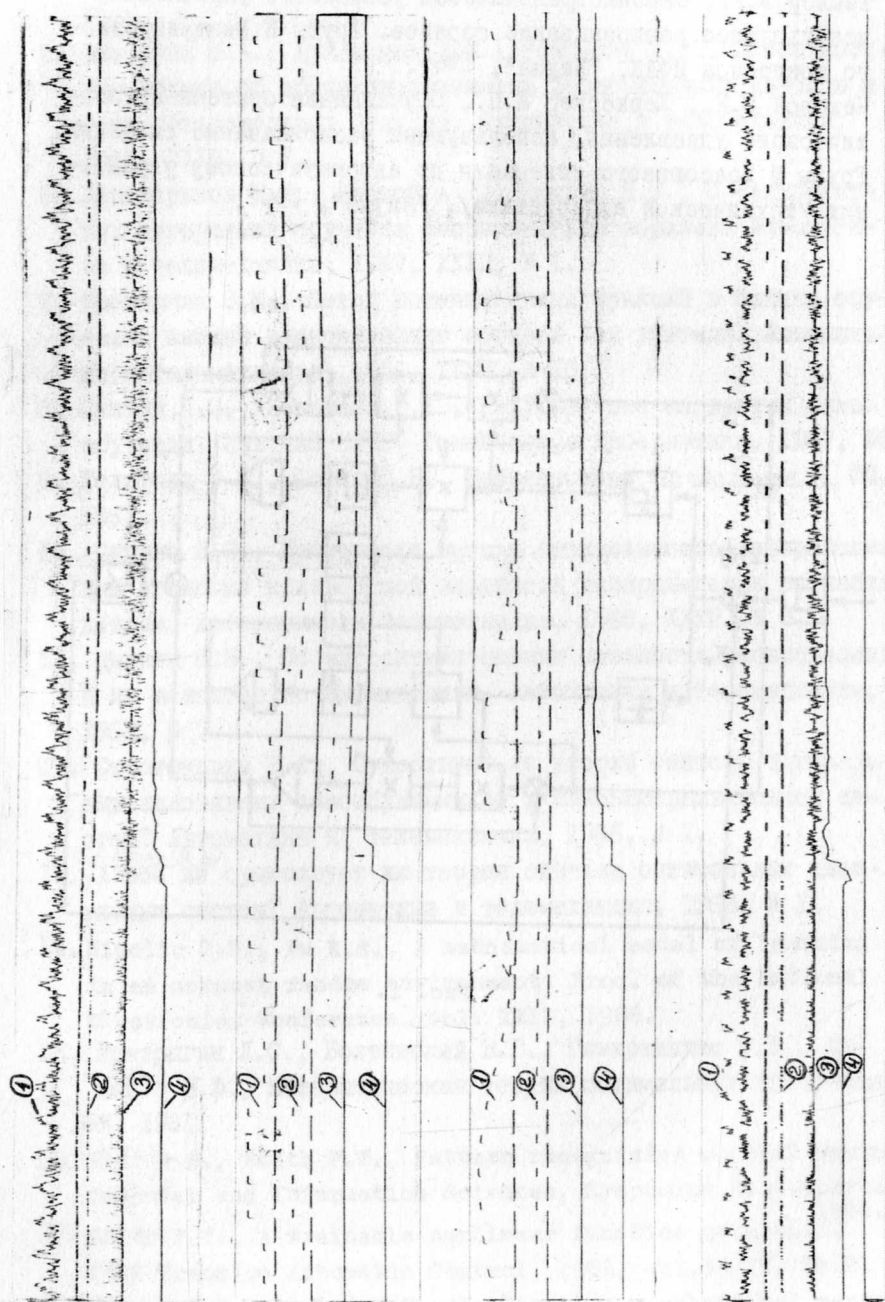


Рис. 1.



DETERMINATION OF THE OPTIMAL MEASUREMENT SEQUENCE FOR A NOISY PATTERN RECOGNITION SYSTEM VIA THE DISCRETE MAXIMUM PRINCIPLE

Marvin H. Hammond, Jr., PhD
Member of the Technical Staff
Missiles Research and Engineering
North American Rockwell Corporation
Columbus, Ohio

Herman R. Weed, Prof.
Department of Electrical
Engineering
Ohio State University
Columbus, Ohio

INTRODUCTION

A basic problem within the pattern recognition specialty area is the determination of the best sequence in which the measurements of an unknown pattern should be taken. This paper presents a method for solving this problem in which the best sequence is defined as the one which minimizes a performance functional which includes the various costs associated with the problem. In particular these costs are the cost incurred as each kind of measurement is taken and the cost resulting from an incorrect classification. The first of these two types of costs is considered to vary as a function of the times when each measurement is taken and the second type is also considered to be time dependent, i.e., premature terminations of a process can occur.

Initial work^{9,10} within this specialty area utilized a concept of a pattern space where each coordinate axis of this space corresponded to a different feature of the pattern. Geometric hypersurfaces were used to divide this space into the various pattern classes and their locations were dependent upon the ranges of the values of each of the features within each pattern class. This type of approach however, did not allow the application of a sequential procedure since no decision could be made until all measurements had been taken. Other researchers^{1,2} have applied the basic sequential decision techniques of Wald¹¹ and Goode⁴ to this type of problem.

The most recent work in the sequential ordering of gathering information was by Fu, Chien, and Cardillo³. This paper proposed the use of dynamic programming in the solution of this basic problem. Their work, however, is valid only for the simple noiseless type problem. They conjectured that modifications existed to their approach which would make it applicable to the more general problem. The present paper includes these modifications in utilizing a form of the discrete Pontryagin's Maximum Principle in the problem solution.⁵

THE DYNAMICAL EQUATION AND THE PERFORMANCE FUNCTIONAL

Development of the Dynamical Equation

The basic components of this problem are the kinds of measurements which can be taken, the values resulting from the measurements, and the various classes in which the unknown pattern could belong. Accordingly the following symbols are defined:

s : a set composed of the various classes, i.e., $s = \{s_1, s_2, \dots, s_\alpha\}$ where s_i is the i^{th} pattern class and there are α such classes.

$u(t)$: the control variable representing the kind of measurement taken at time t , e.g., $u(t) = u_k$ where $k = 1, 2, \dots$, or β .

We further define the set u^t as

$$u^t = \{u(t), u(t-1), \dots, u(1)\} \quad (1)$$

$v(t)$: the value observed when a measurement is taken at time t . The range of these values is assumed to be quantized into γ_k levels where the k refers to the kind of measurement taken. It is assumed that one has available for use in the solution of the problem

$$P(v(t) = v_l \mid u(t) = u_k, s_i) \quad (2)$$

for $l = 1, 2, \dots, \gamma_k$, $k = 1, 2, \dots, \beta$, $i = 1, 2, \dots, \alpha$. Measurement value uncertainty, i.e., noise, enters through this probability. The exact manner and the significance of considering this noise is shown in Appendix A. Similar to the above definition v^t is defined as

$$v^t = \{v(t), v(t-1), \dots, v(1)\} \quad (3)$$

θ : the sum total of information known about the unknown pattern prior to the taking of any measurements.

Using these symbols, it is now possible to express the initial probability that the unknown pattern lies within the i^{th} pattern class as $P(s_i, 0 \mid \theta)$. After $t-1$ measurements have been taken and the corresponding measurement values observed, the prior probability at time t would be $P(s_i, t \mid u^{t-1}, v^{t-1}, \theta)$. Similarly the posterior probability would be $P(s_i, t \mid u^t, v^t, \theta)$ and they are related through Baye's theorem.

$$P(s_i, t \mid u^t, v^t, \theta) = \frac{P(u(t), v(t) \mid s_i, u^{t-1}, v^{t-1}, \theta) P(s_i, t \mid u^{t-1}, v^{t-1}, \theta)}{\sum_{j=1}^{\alpha} P(u(t), v(t) \mid s_j, u^{t-1}, v^{t-1}, \theta) P(s_j, t \mid u^{t-1}, v^{t-1}, \theta)} \quad (4)$$

Through this equation the posterior probability of one class s_i is related to the prior probability of each class s_j , $j = 1, 2, \dots, \alpha$. As such the complete set of posterior probabilities at time t ,

$$P(s, t | u^t, v^t, \theta) = \{P(s_1, t | u^t, v^t, \theta), P(s_2, t | u^t, v^t, \theta), \dots, P(s_\alpha, t | u^t, v^t, \theta)\} \quad (5)$$

is related to the complete set of prior probabilities at time t ,

$$P(s, t | u^{t-1}, v^{t-1}, \theta) = \{P(s_1, t | u^{t-1}, v^{t-1}, \theta), P(s_2, t | u^{t-1}, v^{t-1}, \theta), \dots, P(s_\alpha, t | u^{t-1}, v^{t-1}, \theta)\} \quad (6)$$

By considering the set in (6) as a point within a probability space[†] of priors at time t and the set in (5) as a point in a posterior probability space at time t , equation (4) can be considered as the equation describing the transition from a point in prior space to a point in posterior space caused by taking the measurement $u(t)$ and observing the measurement value $v(t)$.

For any given kind and value of measurement, there would exist in the posterior space a unique point for every given point in the prior space. The opposite statement would not be valid since many points within the prior space could be transformed with the same kind of measurement and measurement value into a single point within the posterior space.

If these prior and posterior probabilities are quantized into δ levels, each α dimensional probability space will be divided into ξ small regions called bins, where

$$\xi = 1 + \sum_{i=0}^{\alpha-2} \left(\frac{1+\delta-1}{1+1} \right) \quad (7)$$

(For a problem containing three pattern classes and each probability quantized into eleven levels, the resulting binned probability space would be as shown in Figure 1). Associated with each of these bins will be a set of α probabilities given by the coordinates of some interior point.

Designating the set of probabilities associated with the j^{th} posterior space bin as $P_j(s, t | u^t, v^t, \theta)$, i.e.,

[†]A probability space is defined such that for any point within that space (1) every coordinate is less than or equal to one and greater than or equal to zero and (2) the sum of all the coordinates is equal to one.

$$P_j(s, t | u^t, v^t, \theta) = (P_j(s_1, t | u^t, v^t, \theta), P_j(s_2, t | u^t, v^t, \theta), \dots P_j(s_\alpha, t | u^t, v^t, \theta)) \quad (8)$$

and the i -th prior space bin as $P_i(s, t | u^{t-1}, v^{t-1}, \theta)$, the action of taking a measurement $u(t)$ and observing a measurement value $v(t)$ would now cause a transition from a point within a prior space bin, $P_i(s, t | u^{t-1}, v^{t-1}, \theta)$, to a point within some posterior space bin, $P_j(s, t | u^t, v^t, \theta)$. Defining $P[P_j(s, t | u^t, v^t, \theta)]$ as the probability that j -th posterior space bin is reached after t measurements, it can be expressed in terms of $P[P_i(s, t | u^{t-1}, v^{t-1}, \theta)]$ by

$$P[P_j(s, t | u^t, v^t, \theta)] = \sum_{i=1}^{\xi} \sum_{k=1}^{\beta} \sum_{\ell=1}^{\gamma_k} P[P_j(s, t | u(t) = u_k, v(t) = v_\ell, u^{t-1}, v^{t-1}, \theta) | u(t) = u_k, v(t) = v_\ell, P_i(s, t | u^{t-1}, v^{t-1}, \theta)]$$

$$P[u(t) = u_k | P_i(s, t | u^{t-1}, v^{t-1}, \theta)] \quad (9)$$

$$P[v(t) = v_\ell | u(t) = u_k, P_i(s, t | u^{t-1}, v^{t-1}, \theta)]$$

$$P[P_i(s, t | u^{t-1}, v^{t-1}, \theta)]$$

The conditional probability

$$P[P_j(s, t | u^t, v^t, \theta) | u(t) = u_k, v(t) = v_\ell, P_i(s, t | u^{t-1}, v^{t-1}, \theta)]$$

used in equation (10) is defined in the following manner:

$$P[P_j(s, t | u^t, v^t, \theta) | u(t) = u_k, v(t) = v_\ell, P_i(s, t | u^{t-1}, v^{t-1}, \theta)] = 1 \quad (10)$$

if for each s_z , $z = 1, 2, \dots, \alpha$

$$P_j(s_z, t | u^t, v^t, \theta) \doteq$$

$$P[u(t) = u_k, v(t) = v_\ell | s_z, u^{t-1}, v^{t-1}, \theta] P_i(s_z, t | u^{t-1}, v^{t-1}, \theta)$$

$$\sum_{w=1}^{\alpha} P[u(t) = u_k, v(t) = v_\ell | s_w, u^{t-1}, v^{t-1}, \theta] P_i(s_w, t | u^{t-1}, v^{t-1}, \theta)$$

i.e., the α values which the right hand side assumes as $z = 1, 2, \dots, \alpha$ lie within the j -th probability bin, and equals zero if the $u(t) = u_k$ and $v(t) = v_\ell$ does not transform the state of the system from the i -th prior bin to the j -th posterior bin.[†] This definition follows from the requirement that Baye's theorem gives the relationship between the prior and posterior probabilities.

[†]Some of the original work in using this expression was developed with the help of Mr. Jay N. Burns, Columbus Division, North American Rockwell Corporation, Columbus, Ohio.

This expression can be simplified by defining the tensor elements

$$m_j^{ikl} = P[P_j(s, t | u(t) = u_k, v(t) = v_l, u^{t-1}, v^{t-1}, \theta) | u(t) = u_k, v(t) = v_l, P_1(s, t | u^{t-1}, v^{t-1}, \theta)] \quad (11)$$

$$q_l^{ik}(t) = P[v(t) = v_l | u(t) = u_k, P_1(s, t | u^{t-1}, v^{t-1}, \theta)] \quad (12)$$

$$r_k^i(t) = P[u(t) = u_k | P_1(s, t | u^{t-1}, v^{t-1}, \theta)] \quad (13)$$

$$x_j(t) = P[P_j(s, t | u^t, v^t, \theta)]$$

$$y_1(t) = P[P_1(s, t | u^{t-1}, v^{t-1}, \theta)] \quad (14)$$

If one assumes that no additional information is obtained between the measurements at times $t-1$ and t , then the prior probability at time t equals the posterior probability at time $t-1$. Using the above definitions and assumption equation (9) becomes

$$x_j(t) = \sum_{i=1}^{\xi} \sum_{k=1}^{\beta} \sum_{l=1}^{\gamma_k} m_j^{ikl} q_l^{ik}(t) r_k^i(t) x_1(t-1) \quad (15)$$

Suppressing the summation signs and letting the mere existence of the same script appearing on different levels imply that a summation is to be taken over all possible values of that script, equation (15) becomes

$$x_j(t) = m_j^{ikl} q_l^{ik}(t) r_k^i(t) x_1(t-1) \quad (16)$$

This dynamical equation describes the changes in the probability that the system is in each of the ξ possible bins or states. The problem has therefore been placed in the format of a stochastic finite state system.

The Performance Functional

The selection procedure of determining the best measurement sequence is to be based upon a minimization of a performance functional which is equal to the expected total cost. For this problem this cost expression would be equal to the sum of the following terms:

1. The sum over all time $t = 1, 2, \dots, t_f$ of the product of the cost of each measurement at time t and the probability that it was taken at time t ,
2. The product of the cost incurred if an incorrect terminal decision is made and the probability that a terminal decision will be made,
3. The sum over $t = 0, 1, 2, \dots, t_f - 1$ of the product of the cost

incurred if a premature decision is made at time t and the probability that an early decision is required at time t .

The numerical value of these incorrect decision costs can be determined from a loss matrix where the lj term of this matrix, $L_{lj}(t)$, equals the cost incurred if the decision placed the unknown pattern in class l when the actual membership was in class j and this occurred at time t . The expected decision cost incurred if the system was in state i at time t , i.e., the i^{th} set of α probabilities has been reached at time t , and the decision was to place the pattern in the l^{th} pattern class would be

$${}_i\mathcal{L}_l(t) = \sum_{j=1}^{\alpha} L_{lj}(t) P_i(s_j, t | u^t, v^t, \theta) \quad (17)$$

Given that the system was in state i at time t , the best decision would be to place the pattern in the class associated with the minimum expected decision cost, i.e., choose the membership to be with the class l^* where

$${}_i\mathcal{L}_{l^*}(t) = \min_l {}_i\mathcal{L}_l(t) \quad (18)$$

Since the set

$$\{P_i(s_1, t | u^t, v^t, \theta) \dots P_i(s_{\alpha}, t | u^t, v^t, \theta)\}$$

has been defined, i.e., the i^{th} bin is composed of a particular set of probability values, the ${}_i\mathcal{L}_{l^*}(t)$ value could be calculated for each bin in the space. The l^* for each bin would then correspond to the pattern class to which the system in state i should assign the membership of the unknown pattern if a decision was to be made.

The value of ${}_i\mathcal{L}_{l^*}(t)$ for $t = t_f$ would be the expected terminal loss due to an incorrect decision given the system was in state i . Multiplying this by a desired probability that a terminal decision is to be made, $p(t_f)$, the variable e^1 can be defined as

$$e^1 = p(t_f) {}_i\mathcal{L}_{l^*}(t_f) \quad (19)$$

${}_i\mathcal{L}_{l^*}(t)$, for $t \neq t_f$, is the expected loss encountered if a premature decision is to be made at time t . If the probability of a desired early termination is given by $p(t)$, $t = 0, 1, \dots, t_f - 1$, then the variable $d^1(t)$ can be defined as

$$d^1(t) = {}_i\mathcal{L}_{l^*}(t) p(t) \quad (20)$$

Using these cost expressions and the symbols defined above to indicate the probability state at each instant of time and the measurement

used, a mathematical expression of the expected total cost would be

$$J = e^i x_i(t_f) + \sum_{t=0}^{t_f-1} [c^k(t+1) r_k^i(t+1) x_i(t) + d^i(t) x_i(t)] \quad (21)$$

EXTENDING THE MAXIMUM PRINCIPLE TO THIS PATTERN RECOGNITION PROBLEM

The Maximum Principle for a Coefficient Type Controller

Following the procedures of Katz⁸ and Jordan and Polak⁶, the Hamiltonian for this problem is

$$H[\lambda(t+1), x(t), r(t+1)] = \lambda^j(t+1) m_j^{ikl} q_l^{ik}(t+1) r_k^i(t+1) x_i(t) + c^k(t+1) r_k^i(t+1) x_i(t) + d^i(t) x_i(t) \quad (22)$$

where $\lambda(t)$ is a ξ dimensional adjoint vector which varies as

$$\lambda^i(t) = \frac{\partial H}{\partial x_i(t)} = \lambda^j(t+1) m_j^{ikl} q_l^{ik}(t+1) r_k^i(t+1) + c^k(t+1) r_k^i(t+1) + d^i(t) \quad (23)$$

The terminal constraints on this vector are

$$\lambda^i(t_f) = e^i \quad i = 1, 2, \dots, \xi \quad (24)$$

These authors have proven that in order to minimize the performance functional in equation (21) subject to the dynamical equation constraint, (16), one must choose those values of $r_k^i(t)$ at each time step such that $H[\lambda(t), x(t-1), r(t)]$ is minimized.

Now Kashap⁷ has proved that the set of allowed values of $r_k^i(t)$ which do satisfy the minimization requirement of the $H[\lambda(t), x(t-1), r(t)]$ are the values zero and one, i.e.

$$r_k^i(t) = \begin{cases} 1 & \text{for } k=k^* \\ 0 & \text{for } k \neq k^* \end{cases} \quad (25)$$

where k^* is the appropriate k which causes $H[\lambda(t), x(t-1), r(t)]$ to be a minimum. This means that the optimal solution agrees with a general constraint that only one measurement is to be taken at any specific time.

The solution of the two ξ dimensional difference equations (16) and (23) is to be based upon a backward calculation procedure and the optimal value of $r_k^i(t) \forall i, k, t$ will be found by choosing those values which minimize the Hamiltonian at each time step. The specific procedure to be used is described in the next subsection.

The Procedures for Determining the Optimal Sequence

By defining the function $g^{ik}(t+1)$ as

$$g^{ik}(t+1) = \lambda^j(t+1) m_j^{ik} q_f^{ik}(t+1) + c^k(t+1) \quad (26)$$

The Hamiltonian becomes

$$H = g^{ik}(t+1) r_k^{i1}(t+1) x_i(t) + d^i(t) x_i(t) \quad (27)$$

Since Kashyap proved that the optimal values of $r_k^{i1}(t+1)$ are zeros and ones, the minimum value of H would be obtained if for each i , $i = 1, 2, \dots$:

$$r_k^{i1}(t+1) = \begin{cases} 1 & \text{if } k = k^* \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

where k^* is defined from

$$\begin{aligned} g^{ik^*}(t+1) &= \min_{k \in \omega_i(t)} g^{ik}(t+1) \\ &= \min_{k \in \omega_i(t)} [\lambda^j(t+1) m_j^{ik} q_f^{ik}(t+1) + c^k(t+1)] \end{aligned} \quad (29)^\dagger$$

and $\omega_i(t)$ is the set of values that k can assume with the system in state i at time t . This $\omega_i(t)$ is included because the values of the m_j^{ik} and q_f^{ik} can change depending upon the nature of the measurements being considered, i.e., is the measurement being taken for the first, second, or n^{th} time.

The basic procedure therefore becomes

1. Initially calculate $\lambda(t_f)$ from

$$\lambda^j(t_f) = e^j \quad (30)$$

2. Iteratively calculate for $t = t_f, t_f-1, \dots, 1$

$$g^{ik^*}(t) = \min_{k \in \omega_i(t)} [\lambda^j(t) m_j^{ik} q_f^{ik}(t) + c^k(t)] \quad (31)$$

$$r_k^{i1}(t) = \begin{cases} 1 & k=k^* \\ 0 & k \neq k^* \end{cases} \quad (32)$$

$$\begin{aligned} \lambda^i(t-1) &= g^{ik^*}(t) + d^i(t) \\ i &= 1, 2, \dots \end{aligned} \quad (33)$$

Since the procedure used in determining the $r_k^{i1}(t)$ values, $\forall i, k, t$, involved a backward calculation, the true set of measurements available

[†]Situations can arise when the specification of just the state of the system would be insufficient to completely specify the set of available k 's. In such cases either previous states in which the system was located or some other specifications can be used to correctly label the set.

at time t , i.e., $\Omega_i(t)$, would have been dependent upon the measurements used at all times t' , $1 \leq t' < t$. As such if one had previously determined that the measurement $u_k \in \Omega_i(t)$ was the one which was the optimal measurement if the system was in state i at time t , the iterative nature of the problem solution would require that a reconsideration be made of this optimal measurement if a measurement u_k was found to be optimal at a time $t' < t$. This reconsideration would be necessary because the set of allowed measurements at t would have to be changed, i.e., $\Omega_i(t)$ would no longer contain u_k but would contain u_k , instead where u_k indicates the next measurement of that feature.

For this example consider the case where u_k taken at time $t' < t$ was the first measurement of the k^{th} feature of the unknown pattern. Then at time t a first measurement can not be taken of that k^{th} feature because a measurement of that feature at t would truly be the second measurement of that feature. For this example the set of measurements allowed in the reconsideration at time t would be labeled $\Omega'_i(t)$ and this revised set would equal the set $\Omega_i(t)$ less the measurement $u_k(t)$ plus the measurement $u_{k'}(t)$.

There are two situations which can arise in this process of recalculation, either $\tilde{k} = k$ or $\tilde{k} \neq k$. For the former case one would need to calculate $g^{\tilde{k}}(t)$ and compare it with the values of $g^{ik}(t)$ determined previously for all the other k 's in $\omega_i(t)$, i.e., for all $k \in \omega_i(t)$, $k \neq \tilde{k}$. The optimal u_k to be used at time t and with the system in state i would be the measurement corresponding to the k^* which minimized $g^{ik}(t)$ with $k \in \omega'_i(t)$. For the case when $\tilde{k} \neq k$ no recalculations would be necessary for the problem. This arises because the total cost incurred when a second measurement is taken is more than or equal that incurred when the first measurement is taken, and since $g^{ik}(t)$ was less than $g^{\tilde{k}}(t)$, by the original minimization criterion, $g^{ik}(t)$ will also be less than $g^{\tilde{k}}(t)$. Therefore a calculation of $g^{\tilde{k}}(t)$ would not be needed nor a second comparison among the $g^{ik}(t)$ for $k \in \omega'_i(t)$; the optimal measurement for the system in state i at time t would remain as u_k . If the problem was changed so that the cost of taking a measurement did vary as to the number of times it had previously been taken, a recalculation of $g^{ik}(t)$ and recomparison among the $g^{ik}(t)$, $k \in \omega'_i(t)$ would always be necessary.

COMPARISON WITH AN EXHAUSTIVE SEARCH TECHNIQUE

The number of basic calculations, i.e., addition and multiplication operations, required to determine the best sequence via the Pontryagin

approach is dependent upon the results of the step by step process. Recalculation of some of the $g^{ik}(t)$'s and $\lambda^i(t)$'s could be required under certain circumstances. As such there would be two extreme cases that could arise, the case where no recalculations would be necessary at any time, and the case where a recalculation would be required at each time one could occur. Defining μ as the probability that a recalculation would be required at any given time one could occur, an upper bound on the total expected number of basic calculations needed to determine the best sequence would be

$$N_{P,\mu} = n\xi(3\beta\gamma_M + \beta + 1) + (1/6)\mu\xi\gamma_M n(n+1)(n-1)(3\gamma_M + \beta + 1) \quad (34)$$

Until the introduction of optimal techniques to aid in the solution of this problem, only an exhaustive search technique was available wherein all possible sequences of measurement would be considered and the one giving the smallest total cost would be chosen as the best sequence. In comparing these two techniques the probability space is assumed to be quantized in the same manner and the same cost expression and transition equation is used. The number of basic calculations required to carry out this procedure would be

$$N_E = \beta^n [2(n+1)(\xi+1)-1 + n\xi(\xi + \gamma_n)] \quad (35)$$

It can be shown that $N_{P,\mu}$ is less than N_E for at least all possible combinations of the following values of the parameters: $1 < \xi < 500$, $2 < \beta < 10$, $2 < n < \beta$, $1 < \gamma_M < 10$. Table 1 gives the values of $N_{P,\mu}$ and N_E for a representative type problem where there are two pattern classes, the probability values are quantized into eleven levels, four kinds of measurements can be taken with a maximum of three possible measurement values observable for each kind of measurement, and a maximum of three measurements allowed to be taken.

CONCLUSIONS

The basic problem of determining a best measurement sequence based upon a minimization of a prescribed cost function has been solved using a discrete version of Pontryagin's Maximum Principle. It has been shown that by using this technique a significantly smaller number of basic arithmetic calculations are required than that needed by a brute force

$$\gamma_M = \max_k \gamma_k, \quad n = \text{number of measurements allowed i.e., } n = t_f$$

approach. Since this basic problem in pattern recognition is common to many other areas within general decision theory, this new technique holds much promise in the savings of time and money in solving for the best selection sequence of gathering information.

Examples of the types of problems in which this technique can be applied are found in a wide range of fields. The decision of which new type of consumer item should be manufactured would need to be based upon an optimal gathering of expensive information about the buying public. The technique described above would prescribe the best sequential order that such information should be obtained. This technique could also be applied in a school testing procedure wherein the correct classification of a "student" can be made with a minimal expenditure of effort. Another direct application of this procedure would be the initial programming of the sequence of measurements to be taken by a device which would recognize patterns.

Table 1
Comparison Between Approaches

Approaches	Number of Arithmetic Calculations
Pontryagin Maximum Principle	
With no recalculations	1,353
With the maximum number of recalculations	3,201
With recalculations occurring with a probability $\mu = 1/\beta$	1,815
Exhaustive Search	36,648
Problem Constants $\alpha = 2$, $\delta = 11$, $\beta = 4$, $\gamma_M = 3$, $n = 3$	

APPENDIX A - NOISE CONSIDERATIONS

The form of the noise considered is a probability, $P(v_i | v_j', u_k)$, over the range of possible measurement values that could be observed given the correct measurement value and the kind of measurement, i.e.,

the noise in the measuring instrument is considered. It is assumed that the probability $P(v_j' | u_k, s_1)$ was previously determined. $P[v(t) = v_j | u(t) = u_k, s_1]$ can therefore be found by

$$P[v(t) = v_j | u(t) = u_k, s_1] = \sum_{j=1}^{\gamma_k} P(v_j | v_j', u_k) P(v_j' | u_k, s_1) \quad (A1)$$

Since real world problems include such noise, a useful decision aiding procedure must consider its effects. One of those effects is the change in the measurement value probabilities as indicated above. However the most significant effect is the fact that repeated measurements can give additional information and therefore the capability of repeated use of all measurements must also be considered.

BIBLIOGRAPHY

1. Birdsall, T. G. and Roberts, R. A., "On the Theory of Signal Detectability, an Optimum Nonsequential Observation-Decision Procedure," IEEE Transactions on Information Theory, Vol. 11, No. 2, April 1965, pp 195-204.
2. Fu, K. S. and Chien, Y. T., "On the Finite Stopping Rules and Non-parametric Techniques in a Feature-Ordered Sequential Recognition System," TR-EE-16; School of Electrical Engineering, Purdue University.
3. Fu, K. S., Chien, Y. T., and Cardillo, G. P. "A Dynamic Programming Approach to Sequential Pattern Recognition" IEEE Transactions on Electronic Computers, Vol. EC-16, No. 6, December 1967, pp 790-803.
4. Goode, H. H., "Deferred Decision Theory," article in Recent Developments in Information and Decision Processes by Robert E. Machol and Paul Gray, The Macmillan Company, New York, 1962.
5. Hammond, Marvin H., Jr. "Optimum Finite Sequential Pattern Recognition Using the Maximum Principle of Pontryagin" North American Rockwell Corporation Report NR68H-203, 3 January 1968, PhD dissertation Ohio State University, March 1968
6. Jordan, B. W. and Polak, E., "Theory of a Class of Discrete Optimal Control Systems," Journal of Electronics and Control, Vol. 17, No. 6, p 697, December 1964
7. Kashyap, R. L., "Optimization of Stochastic Finite State Systems," IEEE Transactions on Automatic Control, Vol. AC-11, No. 95, October 1966, pp 685-692.

8. Katz, Stanley, "A Discrete Version of Pontryagin's Maximum Principle," Journal of Electronics and Control, Vol. 13, No. 2, August 1962.
9. Nillson, Nills J., Learning Machines, McGraw-Hill Book Company, New York, 1965.
10. Sebestyen, G., Decision-Making Process in Pattern Recognition. The MacMillan Company, 1962.
11. Wald, Abraham, Statistical Decision Functions, John Wiley & Sons, Inc., New York, 1950.

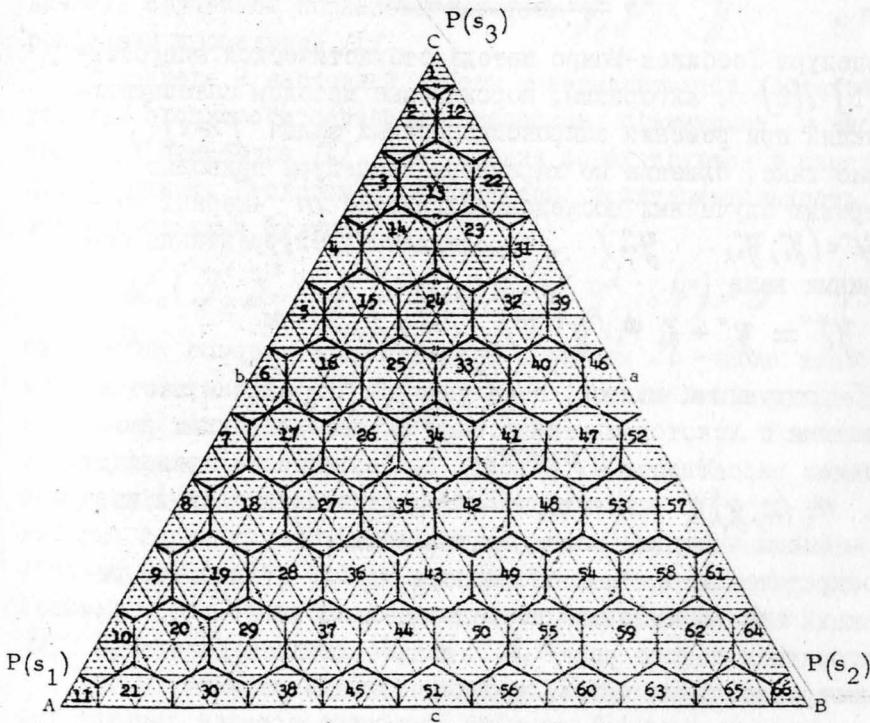


Figure 1. Binned Probability Space for $\alpha = 3$, $\delta = 11$, $\xi = 66$

СХОДИМОСТЬ АЛГОРИТМОВ ОБУЧЕНИЯ И АДАПТАЦИИ

Э.М.Браверман, Б.М.Литваков

ИАТ/ТК/

Москва

СССР

1. Введение

Процедуры Роббинса-Монро метода стохастической аппроксимации [1], [2], алгоритмы, порождаемые методом потенциальных функций при решении аппроксимационных задач [3-7] и некоторые иные, близкие по характеру процедуры приводят к рассмотрению случайных последовательностей m -мерных векторов $y^n = (y_1^n, y_2^n, \dots, y_m^n)$, определяемых рекуррентными соотношениями вида

$$y_i^{n+1} = y_i^n + \gamma_n \phi_i(y^n, x^n), \quad i = 1, 2, \dots, m \quad /1/$$

где x^n - случайный вектор, появляющийся на каждом шаге в соответствии с некоторым неизвестным заранее условным распределением вероятности $P(x^n/y^n)$, явно не зависящим от n , $\phi_i(x, y)$ - некоторые детерминированные функции, а γ_n - члены числовой последовательности.^{x/}

Конкретные алгоритмы, сводящиеся к /1/, отличаются видом функций ϕ_i и числовой последовательности γ_n . Обычно в алгоритмах такого рода γ_n некоторая последовательность неотрицательных чисел, удовлетворяющая условиям

$$\sum \gamma_n = \infty \quad /2a/$$

x/ На связь процедуры Роббинса-Монро с процедурами метода потенциальных функций впервые обратил внимание Я.З.Цыпкин [8]. Он же показал [9], что процедура типа /1/ может быть использована для решения широкого класса задач теории адаптации и обучения машин. Общность методов доказательства сходимости процедуры Роббинса-Монро и процедур метода потенциальных функций была отмечена в [10].

$$\sum \gamma_n^2 < \infty$$

/26/

В некоторых случаях, однако, условие /26/ может быть заменено более слабым условием

$$\lim_{n \rightarrow \infty} \gamma_n = 0$$

/2в/

/см., например, теорему УИ настоящей работы/.

Естественно, возникает задача об установлении критериев сходимости случайной последовательности $y^1, y^2, \dots, y^n$, порождаемой процедурой /I/.

В разделе 2 настоящей работы устанавливаются достаточные условия сходимости случайных процессов, применимые, в частности, к процедуре /I/. Эти условия формулируются в следующих терминах. Вводится в рассмотрение последовательность детерминированных функций

$$u_n(y^1, y^2, \dots, y^n) \geq 0 \text{ и } v_n(y^1, y^2, \dots, y^n) \geq 0 \quad /3/$$

от, вообще говоря, возрастающего с ростом n числа векторных аргументов $y^i = (y_i^1, y_i^2, \dots, y_i^m)$, являющихся реализациями случайного процесса $y^1, y^2, \dots, y^n$. Устанавливаются такие соотношения для функций u_n и v_n , выполняющиеся в силу свойств случайного процесса $y^1, y^2, \dots, y^n$, которые гарантируют стремление к нулю в том или ином смысле /"по вероятности", "почти наверное" или "в среднем"/, по крайней мере, одной из случайных последовательностей

$$u_1, u_2, \dots, u_n \text{ либо } v_1, v_2, \dots, v_n.$$

При использовании полученных условий сходимости в конкретных задачах удается подобрать функции /3/ так, что из сходимости в некотором смысле одной из этих функций к нулю следует сходимость в том же смысле случайного процесса $y^1, y^2, \dots, y^n$. Опыт использования указанных критериев сходимости при рассмотрении конкретных задач /см., например, [4-7] / показывает, что подбор таких функций во многих случаях не вызывает затруднений. Во всех теоремах раздела 2 предполагается, что функции /3/ в силу случайного процесса $y^1, y^2, \dots, y^n$ удовлетворяют следующему условию.

Условие А: Математические ожидания $M\{u_n(y')\}$, $M\{v_n(y')\}$ существуют и

$$M\{u_{n+1}(y', \dots, y'') / y', \dots, y''\} \leq (1 + M_n) \cdot u_n(y', \dots, y'') - \gamma_n v_n(y', \dots, y'') + \xi_n, \quad n = 1, 2, \dots \quad /4/$$

где $\gamma_n \geq 0$ и M_n — последовательности, такие, что

$$a/ \sum_{n=1}^{\infty} \gamma_n = \infty$$

$$б/ \sum_{n=1}^{\infty} |M_n| < \infty$$

а ξ_n — некоторая случайная последовательность.

В формуле /4/ символ $M\{\dots / \dots\}$ означает условное математическое ожидание.

Свойства случайной последовательности ξ_n и числовой последовательности γ_n детализируются в каждой из теорем раздела 2. В приложениях теорем раздела 2 к процедуре /1/ оказывается, что $M\{\xi_n\}$ и $|M_n|$ пропорциональны γ_n^2 .

Условие А вместе с детализацией свойств последовательностей ξ_n и γ_n составляют первое условие всех теорем раздела 2. Эти теоремы различаются, по существу, своими вторыми условиями, устанавливающими дополнительные требования к связи между функциями $u_n(y', \dots, y'')$ и $v_n(y', \dots, y'')$, которые вместе с первыми условиями теорем и позволяют доказывать сходимость соответствующих случайных последовательностей.

Вторые условия теорем раздела 2 удается существенно ослабить, если тем или иным способом установлена ограниченность почти всех реализаций случайного процесса $y', \dots, y''$.

Определение. Последовательность функций $u_n(y', \dots, y'')$ называется бесконечно-большой, если любая последовательность $y', \dots, y'', \dots$, для которой предел $\lim_{n \rightarrow \infty} u_n(y', \dots, y'')$ существует и конечен, ограничена.

Для ряда теорем раздела 2 удается показать, что почти все реализации случайного процесса $y', \dots, y'', \dots$ действительно ограничены, если последовательность функций u_n

не только удовлетворяет первым условиям этих теорем, но и является бесконечно большой. Модифицированные соответствующим образом теоремы обозначаются в тексте работы тем же номером с индексом "а" /например, модификация теоремы I фигурирует в тексте как теорема Ia/.

Предлагаемый в настоящей работе подход к установлению сходимости случайных процессов близок по идее к прямому методу Ляпунова исследования устойчивости движения. В методе Ляпунова факт устойчивости устанавливается, если удастся подобрать некоторую функцию фазовых координат, удовлетворяющую условиям, обеспечивающим стремление ее к нулю в процессе возмущенного движения. Как видно из условия А, функции U_n и $Y_n V_n$ играют роль "функции Ляпунова" и "ее производной в силу процесса".

Сама по себе идея применения аналогов метода Ляпунова для исследования сходимости стохастических процессов использовалась в целом ряде работ /см., например, [11 - 13] /. Настоящая работа отличается тем, что условия, накладываемые на "функцию Ляпунова", отражают специфику случайных процессов, порождаемых соотношениями /1/ и /2/ ^{х/}. Известные теоремы Блума [2], используемые для доказательства сходимости процедур метода стохастической аппроксимации имеют такой же характер.

В разделе 3 общие теоремы предыдущего раздела используются непосредственно для установления условий сходимости процедуры /1/. Здесь формулируется теорема У, условия которой по своей форме аналогичны теоремам Блума [2]. Эта теорема, из которой результаты Блума и Е.Г.Гладышева [15] следуют

х/ Непрерывным детерминированным аналогом процесса /1/,
/2/ служат уравнения

$$\frac{dy_i}{dt} = \gamma(t) \phi_i(y_1, \dots, y_m); \quad i = 1, \dots, m$$

при условии

$$\int_0^{\infty} \gamma(t) dt = \infty$$

$$\int_0^{\infty} \gamma^2(t) dt < \infty$$

$$/или \lim_{t \rightarrow \infty} \gamma(t) = 0/$$

как частный случай, позволяют устанавливать сходимость процедуры /I/, применяемой для решения уравнений регрессии и тогда, когда нельзя гарантировать единственность решения этих уравнений. Следующая теорема УІ позволяет устанавливать сходимость процедуры /I/, исходя из иных свойств порождаемого процедурой /I/ процесса, нежели тех, которые используются в теореме УІ. Кроме того, в заключение раздела 3 приводится теорема УІІ, распространяющая результаты Блума и на тот случай, когда последовательность χ_n , фигурирующая в процедуре /I/, удовлетворяет не условию /2б/, а более слабому условию /2в/.

2. Основные теоремы о сходимости

Рассмотрим случайный процесс в дискретном времени n , определяемый условными плотностями вероятности $P_{n+1}(y^{n+1}/y^1, \dots, y^n)$ появления случайной величины y^{n+1} в $n+1$ -ый момент времени при условии, что в моменты $1, \dots, n$ случайные величины принимали значения $y^1, y^2, \dots, y^n$ соответственно.

Достаточные условия сходимости к нулю математических ожиданий $M\{u_n\}$ и $M\{v_n\}$ последовательности функций /З/, заданных на случайном процессе $y^1, \dots, y^n$, устанавливаются теоремами I и II. В силу неотрицательности величин u_n и v_n отсюда следует очевидным образом и сходимость этих величин к нулю по вероятности.

Теорема I. Пусть задан случайный процесс $y^1, \dots, y^n$ и последовательность скалярных функций /З/, удовлетворяющих условиям :

1⁰. условию А, причем

$$\lim_{n \rightarrow \infty} \chi_n = 0; \quad \sum_1^\infty M\{\chi_n\} < \infty$$

$$2^0. \quad M\{v_{n+1}/y^1, \dots, y^n\} \leq (1 + A\chi_n) \cdot v_n(y^1, \dots, y^n) + B\chi_n + \zeta_n$$

где A и B - некоторые константы, а $\zeta_n \geq 0$ - такая случайная последовательность, что

$$\sum_1^\infty M\{\zeta_n\} < \infty$$

Тогда

$$\lim_{n \rightarrow \infty} M\{V_n\} = 0$$

/5/

и последовательность случайных величин V_n стремится к нулю по вероятности при $n \rightarrow \infty$.

При исследовании конкретных процессов оказывается, что величины A и B , фигурирующие в условии 2° теоремы I, являются не константами, а некоторыми функциями $y', \dots, y''$:

$$A = A_n(y', \dots, y''), \quad B = B_n(y', \dots, y'')$$

Однако, если удастся из каких-либо соображений установить ограниченность последовательностей A_n и B_n для почти всех реализаций случайного процесса, то сходимости V_n по вероятности к нулю может быть доказана и в этом случае. Ограниченность же последовательностей A_n и B_n может следовать, в частности, из ограниченности реализаций случайного процесса. Условие, гарантирующее ограниченность почти всех реализаций случайного процесса $y', \dots, y''$, как уже указывалось в разделе I, могут быть выражены в форме ограничения на вид последовательности функций

$U_n(y', \dots, y'')$. Именно, достаточно потребовать, чтобы последовательность U_n была бесконечно большой.

Теорема Ia: Пусть функции /3/ удовлетворяют условиям: 1°. условию A, причем

$$\lim_{n \rightarrow \infty} \gamma_n = 0; \quad \sum_i M\{\zeta_n\} < \infty$$

$$2^\circ. M\{V_n / y', \dots, y''\} \leq \{1 + A_n(y', \dots, y'')\} \cdot V_n + B_n(y', \dots, y'') \cdot \gamma_n + \zeta_n$$

где $A_n(y', \dots, y'')$ и $B_n(y', \dots, y'')$ - функции, которые для любой ограниченной последовательности $y', \dots, y''$, $|y'_i| \leq R (i=1, \dots, n)$ мажорируются константами $A(R)$ и $B(R)$ соответственно, а ζ_n - случайная последовательность, такая, что

$$\sum M\{\zeta_n\} < \infty$$

3°. Последовательность функций $U_n(y', \dots, y'')$ - бесконечно-

чно большая.

Тогда последовательность случайных величин V_n стремится к нулю по вероятности при $n \rightarrow \infty$

$$V_n \xrightarrow{P} 0$$

В формулировке теоремы I требование сходимости ряда $\sum M\{z_n\}$ является существенным. Формулируемая ниже теорема II позволяет ослабить это требование за счет существенного усиления условия 2⁰ теоремы I. Отказ в теореме II от требования сходимости ряда $\sum M(z_n)$ позволит в дальнейшем /см. теорему УП/ отказаться в некоторых случаях от требования /2б/ в алгоритмах стохастической аппроксимации, заменив его требованием /2в/.

В отличие от теоремы I в теореме II утверждается стремление к нулю последовательности U_n , а не V_n .

Теорема II. Пусть функции /5/ удовлетворяют условиям: I⁰. условию A, причем

$$\lim_{n \rightarrow \infty} \gamma_n = 0 \text{ и } M\{z_n\} = \alpha_n \gamma_n, \lim_{n \rightarrow \infty} \alpha_n = 0$$

2⁰. существует такая константа $A > 0$, что

$$V_n(y^1, \dots, y^n) \geq A U_n(y^1, \dots, y^n)$$

тогда

$$\lim_{n \rightarrow \infty} M\{U_n\} = 0$$

и последовательность случайных величин U_n стремится к нулю по вероятности при $n \rightarrow \infty$.

Перейдем теперь к установлению достаточных условий сходимости "почти наверное". Сформулированные ниже теоремы III, III а и IV показывают, что некоторая модификация условий теорем I и II позволяет установить не только сходимость по вероятности, но и "почти наверное". При этом, однако, в условиях теоремы III, III а гарантируется сходимость математических ожиданий $M\{U_n^\beta\}$ лишь при $\beta < 1$, а в условиях теорем IV, IV а, сходимость математических ожиданий вообще не гарантируется.

Теорема III. Пусть функции /5/ удовлетворяют условиям:

1⁰. условию А, причем

$$\lim_{n \rightarrow \infty} \gamma_n = 0 \quad \sum_1^{\infty} M\{\tau_n\} < \infty$$

2⁰. на почти всех реализациях случайного процесса, для которых $\lim_{n \rightarrow \infty} V_n = 0$ имеет место $\lim_{n \rightarrow \infty} U_n = 0$

Тогда при $n \rightarrow \infty$ последовательность случайных величин стремится к нулю почти наверное:

$$U_n \xrightarrow{n.в.} 0$$

и, кроме того, при любом $\beta < 1$

$$\lim_{n \rightarrow \infty} M\{U_n^\beta\} = 0$$

/6/

Теорема IIIa. Пусть функции /5/ удовлетворяют условиям:

1⁰. условию А, причем

$$\lim_{n \rightarrow \infty} \gamma_n = 0, \quad \sum_1^{\infty} M\{\tau_n\} < \infty$$

2⁰. на всех ограниченных последовательностях, для которых $\lim_{n \rightarrow \infty} V_n = 0$ имеет место $\lim_{n \rightarrow \infty} U_n = 0$

3⁰. последовательность функций $U_n(y', \dots, y^n)$ - бесконечно большая.

Тогда при $n \rightarrow \infty$ последовательность случайных величин стремится к нулю почти наверное и, кроме того, при любом $\beta < 1$

$$\lim_{n \rightarrow \infty} M\{U_n^\beta\} = 0$$

Используя теорему III, можно доказать сходимость известного процесса А.Дворецкого [14]. При этом установление того факта, что можно выбрать функции U_n и V_n , удовлетворяющие условиям теоремы III, не связано с анализом рассматриваемого случайного процесса, а требует лишь алгебраических преобразований. Анализ же случайного процесса целиком "берет на себя" теорема III. В теореме Дворецкого, рассматривается случайный процесс $x', \dots, x^n, \dots$ и доказывается его сходимость как почти наверное, так и в среднем квадрате, т.е.

$$x^n \xrightarrow{n \rightarrow \infty} 0, \lim_{n \rightarrow \infty} M\{|x^n|^p\} = 0 \quad \text{при } p \leq 2$$

Применение теоремы III позволяет доказать несколько более слабое утверждение

$$x^n \xrightarrow{n \rightarrow \infty} 0, \lim_{n \rightarrow \infty} M\{|x^n|^p\} = 0 \quad \text{при } p < 2$$

/случай $p = 2$ соответствует $\beta = 1$ в формуле /6/ /.

Следующая теорема IV весьма близка по своим условиям к теореме I, но, в отличие от нее, содержит требование об ограничении роста последовательности V_n /условия 2°/ не в смысле математических ожиданий, а на почти каждой реализации. Это и позволяет установить сходимость к нулю последовательности V_n почти наверное.

Теорема IV. Пусть функции /3/ удовлетворяют условиям:

1°. условию A, причем

$$\lim_{n \rightarrow \infty} \gamma_n = 0 \quad \sum_1^{\infty} M\{\xi_n\} < \infty$$

2°. для любого $\delta > 0$ найдется такое множество реализаций случайного процесса $y^1, \dots, y^n, \dots$, вероятность которого больше $1 - \delta$ и найдутся такие константы $A_\delta > 0$ и $B_\delta > 0$, что на этом множестве реализаций выполнены неравенства

$$V_{n+1} \leq (1 + A_\delta \gamma_n) V_n + B_\delta \gamma_n + \eta_n \quad /7/$$

где η_n — последовательность таких чисел, что $\sum_1^{\infty} \eta_n$ сходится. Тогда при $n \rightarrow \infty$ последовательность случайных величин V_n стремится к нулю почти наверное.

3. Условия сходимости процедуры Роббинса-Монро метода стохастической аппроксимации

В настоящем разделе результаты предыдущего раздела используются для установления достаточных условий сходимости процедуры Роббинса-Монро. Как известно, эта процедура заключается в построении последовательности векторов $y^1, \dots, y^n, \dots$, удовлетворяющей соотношению /I/, и предназначена для ре-

шения системы уравнений регрессии

$$M_x \{ \varphi_i(x, y_1, y_2, \dots, y_m) \} = 0 \quad i=1, \dots, m, /9/$$

где x - случайная величина с фиксированной, но неизвестной функцией $P(x/y)$ распределения вероятности /быть может, зависящей от $y=(y_1, \dots, y_m)$, а символ $M_x \{ \dots \}$ означает математическое ожидание при каждом фиксированном векторе $y=(y_1, \dots, y_m)$. При этом возникает задача установления сходимости в том или ином смысле случайной последовательности $y_1^*, \dots, y_n^*$ к корням y^* системы уравнений /9/.

Приводимые ниже теоремы У - УП обобщают и дополняют известные результаты /см. библиографию в [17]/, устанавливающие условия сходимости. Теорема У охватывает результаты Блума [2] и Е.Г.Гладышева [15], распространяя их, в частности, на тот случай, когда решение уравнений регрессии может быть и не единственным.

Теорема УI также устанавливает условия сходимости процедуры /I/ в случае, когда единственность решения системы уравнений /9/ не может быть гарантирована. Эта теорема имеет несколько иной характер, нежели теорема У и не может быть получена из нее.

Теорема УП модифицирует условия теоремы У для случая, когда последовательность y_n , фигурирующая в /-/ , не удовлетворяет условию /26/.

Следуя Блуму, введем в рассмотрение неотрицательную, дважды непрерывно дифференцируемую функцию $u(y)$ и функции

$$V(y) \equiv - M_x \left\{ \sum_{i=1}^m \frac{\partial u}{\partial y_i} \varphi_i(x, y) \right\}$$

$$W_n(y) \equiv M_x \left\{ \max_{0 \leq \theta \leq 1} \sum_{i,k=1}^m \left(\frac{\partial^2 u / \partial z_i \partial z_k}{\partial z_i \partial z_k} \right)_{z=y+\theta y_n} \cdot \varphi_i(x, y) \varphi_k(x, y) \right\}$$

которые предполагаются существующими. Во всех теоремах на-

стоящего раздела предполагается, что выполнено следующее условие.

Условие Б. $V(y) \geq 0$

и при любом $n \geq 1$

$$W_n(y) \leq a U(y) + b V(y) + c$$

где a, b, c — некоторые константы.

Отметим основные моменты использования теорем предыдущего раздела для доказательства теорем УІ–УІІ. Роль условия Б в теоремах настоящего раздела заключается в том, что выполнение этого условия гарантирует, что случайные последовательности

$$U_n \equiv U(y^n), \quad V_n \equiv V(y^n)$$

в силу рекуррентной процедуры /І/ и соотношений /2а/, /2б/ удовлетворяют условию А раздела І. Для того, чтобы можно было использовать теоремы предыдущего раздела, надо еще обеспечить выполнение их остальных условий. Кроме того, надо гарантировать, что из факта стремления $U(y^n)$ или $V(y^n)$ /к нулю/ а этот факт и устанавливается теоремами предыдущего раздела/, следует сходимость случайной последовательности $y^1, \dots, y^n, \dots$ к решению системы уравнений /9/. Этим целям и служат остальные условия теорем У–УІІ.

Для того, чтобы сформулировать теорему У, обозначим через Y^* множество решений системы уравнений /9/ и будем говорить, что y^n стремится к Y^* при $n \rightarrow \infty$ по вероятности /почти наверное/

$$y^n \xrightarrow{P} Y^* \quad (y^n \xrightarrow{n, P} Y^*)$$

если

$$\rho(y^n, Y^*) \equiv \inf_{y \in Y^*} |y^n - y| \xrightarrow{P} 0 \quad (\rho(y^n, Y^*) \xrightarrow{n, P} 0)$$

Теорема У. Пусть $y^1, \dots, y^n, \dots$ — случайный процесс, определяемый соотношениями /І/ и /2-а/, /2б/. Пусть, далее, функция $U(y)$ удовлетворяет условиям:

1°. условию Б;

2°. для каждой последовательности $y^1, \dots, y^n, \dots$, на которой $\lim_{n \rightarrow \infty} U(y^n) = 0$ одновременно и $\lim_{n \rightarrow \infty} \rho(y^n, Y^*) = 0$;

3°. Для каждой последовательности $y^1, \dots, y^n, \dots$, на которой $\lim_{n \rightarrow \infty} V(y^n) = 0$, одновременно и $\lim_{n \rightarrow \infty} U(y^n) = 0$. Тогда при $n \rightarrow \infty$ случайный вектор y^n стремится к Y^* почти наверное

$$y^n \xrightarrow{n.к.} Y^*$$

Если же функция $U(y)$ бесконечно большая, то условия 2° и 3° могут быть заменены более слабыми условиями:

2а°. Функция $U(y)$ может обращаться в нуль лишь в точках из Y^* .

3а°. Для каждой ограниченной последовательности, на которой $\lim_{n \rightarrow \infty} V(y^n) = 0$, одновременно и $\lim_{n \rightarrow \infty} U(y^n) = 0$.

Доказательство теоремы У опирается на теоремы III и IIIа предыдущего раздела.

Теорема VI. Пусть $y^1, \dots, y^n, \dots$ — случайный процесс, определяемый соотношениями /1/ и /2а/, /2б/. Пусть, далее, функция $U(y)$ удовлетворяет условиям:

1°. условию Б;

2°. $U(y)$ — бесконечно большая функция;

3°. функция $V(y)$ непрерывно дифференцируема и может обращаться в нуль лишь в точках из Y^* ;

4°. функция

$$B_n(y) = M_x \left\{ \max_{0 \leq \theta \leq 1} \sum_{i=1}^m \left(\frac{\partial V(z_1, \dots, z_m)}{\partial z_i} \right) \Phi_i(x, y) \right\} \\ z = y + y_n \theta \Phi(x, y)$$

мажорируется функцией $B(y)$, не зависящей от n , ограниченной в любой ограниченной области изменения переменной $y = (y^1, \dots, y^m)$. Тогда при $n \rightarrow \infty$ случайный вектор y^n стремится к Y^* по вероятности:

$$y^n \xrightarrow{P} Y^*$$

Доказательство теоремы УП опирается на теорему 1а предыдущего раздела.

Теорема УП. Пусть $y^1, \dots, y^n, \dots$ — случайный процесс, определяемый соотношениями /1/ и /2а/, /2в/. Пусть, далее, функция $u(y)$ удовлетворяет условиям:

1°. условию Б;

2°. условию 2° теоремы У1;

3°. $V(y) \geq A u(y)$, где $A > 0$ — некоторая константа.

Тогда при $n \rightarrow \infty$ случайный вектор y^n стремится к Y^* по вероятности

$$y^n \xrightarrow{P} Y^*$$

Условие 3° теоремы УП, очевидно, является существенно более сильным, нежели соответствующее условие теоремы J. Именно благодаря усилению этого условия оказывается возможным отказаться от требования /2б/, заменив его более слабым требованием /2в/. Однако, при этом сходимость случайного процесса $y^1, \dots, y^n, \dots$ доказывается не "почти наверное", а лишь по вероятности. Доказательство теоремы УП опирается на теорему П предыдущего раздела.

Несмотря на то, что условие 3° теоремы УП является достаточно жестким, можно отметить целый ряд случаев, когда в приложениях это условие выполняется. В качестве примера можно привести процедуру Роббинса-Монро для определения среднего значения случайной величины x . При этом

$$y^{n+1} = y^n + \gamma_n (x^n - y^n)$$

и, положив,

$$u(y) = V(y) = (y - M\{x\})^2$$

из теоремы УП заключаем, что $y^n \xrightarrow{P} M\{x\}$, если последовательность γ_n удовлетворяет требованиям /2а/, /2в/.

Условие 3° теоремы УП выполняется также при использовании метода потенциальных функций в некоторых задачах обучения /см. 18 /.

ЛИТЕРАТУРА

- I. H. Robbins, S. Monroe, A stochastic approximation method, *Annals of Math. Stat.*, vol. 22(1951), pp. 400-407.
2. R. Blum, Multidimensional stochastic approximation methods, *Annals of Math. Stat.*, vol. 25(1954), pp. 737-744.
3. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. *Автоматика и телемеханика*, том XXV, № 6, 1964.
4. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Вероятностная задача об обучении автоматов распознаванию классов и метод потенциальных функций. *Автоматика и телемеханика*, т. XXV, № 9, 1964.
5. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в задаче о восстановлении характеристики функционального преобразователя по случайно наблюдаемым точкам. *Автоматика и телемеханика*, т. XXV, № 12, 1964.
6. Браверман Э.М. О методе потенциальных функций. *Автоматика и телемеханика*, т. XXVI, № 12, 1965.
7. Браверман Э.М. Метод потенциальных функций в задаче обучения машины распознаванию образов без учителя. *Автоматика и телемеханика*, № 10, 1966.
8. Цыпкин Я.З. О восстановлении характеристики функционального преобразователя по случайно наблюдаемым точкам. *Автоматика и телемеханика*, XXVI, № 11, 1965.
9. Цыпкин Я.З. Адаптация, обучение и самообучение в автоматических системах. *Автоматика и телемеханика*, № 1, 1966.
10. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Процесс Роббинса-Монро и метод потенциальных функций. *Автоматика и телемеханика*, т. XXVI, № 11, 1965.
11. Кац И.А., Красовский Н.Н. Об устойчивости систем со случайными параметрами. *ПММ*, 1960, т. 24, вып. 5.
12. Kushner H.J., On the stability of stochastic dinamica systems. *Proc. of Nat. Acad. of Science*, vol. 53, No. 1, 1965

13. Kushner H.J. On the construction Liapunov function. ASEE Trans. on Automatic Control, v. AC-10, No. 4, October 1965.
14. Dvoretzky A. On stochastic approximation. Proc. of the 3rd Berkeley Symposium on Mathematical Statistics and Probability 1, 1956.
15. Гладышев Е.Г. О стохастической аппроксимации. Теория вероятностей и ее применение. т.10, № 2, 1965.
16. Литваков Б.М. О сходимости рекуррентных алгоритмов обучения распознаванию образов. Автоматика и телемеханика, № 1, 1968.
17. Логинов Н.В. Методы стохастической аппроксимации /обзор/. Автоматика и телемеханика, № 4, 1966.

AN ALGORITHM FOR CALCULATING RECOGNITION ERRORS IN PATTERN RECOGNITION

by

Keinosuke Fukunaga and Thomas F. Krile
School of Electrical Engineering
Purdue University
Lafayette, Indiana, U.S.A.

I. Introduction

The recognition rate of an optimum Bayes' classifier is an important parameter for both pattern classification and feature selection. Unfortunately, this parameter is very difficult to calculate explicitly, especially when the distributions of the pattern classes have different covariance matrices and high-dimensionality.

This paper presents an algorithm with which to calculate the recognition error rate exactly, where the pattern statistics can have unequal covariance matrices and different a priori class probabilities. An approximation algorithm for calculating recognition error is also developed which leads to a reduction in calculations while maintaining good accuracy.

To illustrate an application of these algorithms, the relationships between the number of pattern vector components used for classification and the recognition rate are examined for the divergence and Bhattacharyya criteria, using data first presented by Marill and Green¹.

II. Recognition Errors

The discussion of this paper is based on the following assumptions:

- (1) Two pattern classes are being separated,
- (2) Pattern vectors from each class form an n -dimensional Gaussian distribution,
- (3) Classification is made by an optimum Bayes' classifier.

Considering two pattern classes to be separated, recognition errors occur in one of two ways: either a pattern vector from class 1 is assigned to class 2, or a pattern vector from class 2 is assigned to class 1. Let $P(\omega_k)$ be the a priori probability of class ω_k , and $p(\tilde{X}/\omega_k)$ be the conditional probability density of pattern vector $\tilde{X}$ when it comes from class ω_k . Applying the pattern vectors to an optimum Bayes' classifier, the first error case can be written as

$$e_1 = \text{Prob}\{P(\omega_1)p(\tilde{X}/\omega_1) < P(\omega_2)p(\tilde{X}/\omega_2)/\tilde{X}\epsilon\omega_1\}P(\omega_1), \quad (1)$$

and the second error case as

$$e_2 = \text{Prob}\{P(\omega_1)p(\tilde{X}/\omega_1) > P(\omega_2)p(\tilde{X}/\omega_2)/\tilde{X}\epsilon\omega_2\}P(\omega_2). \quad (2)$$

The total recognition error is then

$$e = e_1 + e_2. \quad (3)$$

Assuming a multivariate Gaussian distribution for each pattern class, we have

$$p(\tilde{X}/\omega_k) = N(\tilde{m}_k, \tilde{\Sigma}_k). \quad (4)$$

Without losing generality, we

assign the origin of the coordinate system such that

$$\underline{m}_1 = \underline{0} \text{ and } \underline{m}_2 = \underline{m}, \quad (5)$$

where $\underline{m}$ is the difference-of-means vector for the two classes being considered.

To eliminate the problem of working in a multidimensional space, a new one-dimensional variable is introduced as

$$U(\underline{X}) = \underline{X}^T \underline{\Sigma}_1^{-1} \underline{X} - (\underline{X} - \underline{m})^T \underline{\Sigma}_2^{-1} (\underline{X} - \underline{m}) + C, \quad (6)$$

where

$$C = \ln[|\underline{\Sigma}_1|/|\underline{\Sigma}_2|] - 2 \ln[P(w_1)/P(w_2)]. \quad (7)$$

Note that $U(\underline{X})$ is nothing more than the quadratic discriminant function associated with a Bayes' classifier. For the rest of this paper, let $U_k = U(\underline{X})$ when $\underline{X}$ belongs to class w_k .

Now the recognition error for the first and second cases can be written simply as

$$e_1 = \text{Prob}\{U_1 > 0\} P(w_1), \quad (8)$$

and

$$e_2 = \text{Prob}\{U_2 < 0\} P(w_2), \quad (9)$$

where only one-dimensional probability distributions are involved.

III. Characteristic Function of U_k

Since the recognition rate is now expressed in terms of the one-dimensional variable U_k , it can be readily calculated if the probability distribution of U_k is obtained. Applying a suitable transformation, U_k can be expressed in terms of new independent variables. Then it becomes easy to calculate the characteristic function of U_k , which includes all information

concerning the distribution of U_k . Finally, a process of inverse transformations is used to obtain the density, and, ultimately, the distribution of U_k , which is the desired goal.

First let us focus our attention on obtaining the independent components and characteristic function of $U_1: (X \in w_1)$. In order to transform the pattern vector $\underline{X}$ into a new vector whose components are independent, consider the $n \times n$ transformation matrix $\underline{A}$ which satisfies the following relations²:

$$\underline{A} \underline{\Sigma}_1 \underline{A}^T = \underline{I} \text{ and } \underline{A} \underline{\Sigma}_2 \underline{A}^T = \underline{\Lambda} \quad (10)$$

$\underline{\Lambda}$ is a diagonal matrix whose diagonal terms, $\lambda_1, \lambda_2, \dots, \lambda_n$, are the eigenvalues obtained from the determinantal equation

$$|\underline{\Sigma}_2^{-1} \underline{\Lambda} \underline{\Sigma}_1 - \underline{I}| = 0. \quad (11)$$

Thus the i^{th} row of $\underline{A}$ is a normalized eigenvector corresponding to the i^{th} eigenvalue, λ_i , where normalization is carried out to make $\underline{A} \underline{\Sigma}_1 \underline{A}^T$ a unit matrix, $\underline{I}$.

The transformation matrix $\underline{A}$ gives rise to a new variable

$$\underline{Y} = \underline{A} \underline{X}, \quad (12)$$

and a transformed difference-of-means vector

$$\underline{\mu} = \underline{A} \underline{m}. \quad (13)$$

Upon substitution into Eq. (6) with

$\underline{X} \in w_1$, U_1 can be written as a function of the new variable $\underline{Y}$ as

$$\begin{aligned} U_1(\underline{Y}) &= \underline{Y}^T \underline{Y} - (\underline{Y} - \underline{\mu})^T \underline{\Lambda}^{-1} (\underline{Y} - \underline{\mu}) + C \\ &= \sum_{i=1}^n \left\{ y_i^2 \frac{1}{\lambda_i} (y_i - \mu_i)^2 - 2 \ln[P(w_1)/P(w_2)] \right\} - 2 \ln[P(w_1)/P(w_2)] \\ &= \sum_{i=1}^n \left\{ \left(1 - \frac{1}{\lambda_i}\right) (y_i - \frac{\mu_i}{\lambda_i - 1})^2 - \left(\frac{\mu_i}{\lambda_i - 1} + \ln \lambda_i\right) \right\} - 2 \ln[P(w_1)/P(w_2)]. \end{aligned} \quad (14)$$

Now we have U_1 in terms of independent Gaussian random variables y_i , each of which has zero mean and unit variance with respect to class ω_1 .

Remembering that $E(X) = 0$ for $X \in \omega_1$, from Eq. (5) the characteristic function of $U_1(X)$ can be written as

$$M_1(j\omega) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \exp[U_1(X)j\omega] N(X, 0, \Sigma_1) dx_1, \dots, dx_n. \quad (15)$$

Using Eq. (14), $M_1(j\omega)$ can be expressed in terms of the independent variables

$$y_i: \\ M_1(j\omega) = K(j\omega) \prod_{i=1}^n \int_{-\infty}^{\infty} \exp \left[j\omega \left\{ y_i^2 - \frac{1}{\lambda_i} (y_i - \mu_i)^2 - m\lambda_i \right\} \right] \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{y_i^2}{2} \right] dy_i \quad (16)$$

where

$$K(j\omega) = \exp \left\{ -2j\omega m [P(\omega_1)/P(\omega_2)] \right\} \quad (17)$$

By manipulating the integrand into a constant times a Gaussian density, the integration can be easily performed to yield*

$$M_1(j\omega) = K(j\omega) \prod_{i=1}^n F_{1i}(j\omega), \quad (18)$$

where

$$F_{1i}(j\omega) = \frac{1}{\sqrt{1 - j2a_{1i}\omega}} \exp \left[\frac{-(2a_{1i}b_{1i}\omega)^2}{2(1 - j2a_{1i}\omega)} - j\omega h_{1i} \right], \quad (19)$$

and

$$a_{1i} = 1 - \frac{1}{\lambda_i}, \quad (20)$$

$$b_{1i} = \frac{\mu_i}{\lambda_i - 1}, \quad (21)$$

$$h_{1i} = \frac{(a_{1i}b_{1i})^2}{1 - a_{1i}} + m\lambda_i. \quad (22)$$

Or, expressing Eqs. (18) and (19) in terms of a magnitude and angle,

$$M_1(j\omega) = \left\{ \prod_{i=1}^n |F_{1i}(j\omega)| \right\} \exp \left[j \sum_{i=1}^n \angle F_{1i}(j\omega) \right], \quad (23)$$

where

$$|F_{1i}(j\omega)| = [1 + (2a_{1i}\omega)^2]^{-\frac{1}{4}} \exp \left[\frac{-2(a_{1i}b_{1i}\omega)^2}{1 + (2a_{1i}\omega)^2} \right], \quad (24)$$

and

$$\angle F_{1i}(j\omega) = \frac{-2\omega}{n} \left[m P(\omega_1)/P(\omega_2) \right] + \frac{1}{2} \tan^{-1} 2a_{1i}\omega - \omega \left[h_{1i} + \frac{a_{1i}(2a_{1i}b_{1i}\omega)^2}{1 + (2a_{1i}\omega)^2} \right]. \quad (25)$$

Eqs. (18) and (23) indicate that the total characteristic function is the product of n independent characteristic functions, which have the same functional form and differ only in the values of two parameters, λ_i and μ_i . Therefore, the calculations for F_{1i} and $\angle F_{1i}$ are systematic and simple to carry out.

The characteristic function of U_2 for $X \in \omega_2$ can be obtained by a procedure similar to the one used for U_1 . This time, a new transformed variable, Z , and a transformed difference-of-means vector, $\underline{v}$, are defined as

$$Z = (\underline{\Lambda}^{-\frac{1}{2}} \underline{A}) (X - m), \text{ and} \\ \underline{v} = (\underline{\Lambda}^{-\frac{1}{2}} \underline{A}) m = \underline{\Lambda}^{-\frac{1}{2}} \underline{\mu}. \quad (26)$$

$\underline{\Lambda}$ and $\underline{A}$, which are the same matrices as defined for the U_1 case, satisfy the following relationships:

$$(\underline{\Lambda}^{-\frac{1}{2}} \underline{A}) \sum_{i=1}^n (\underline{\Lambda}^{-\frac{1}{2}} \underline{A})^T = \underline{\Lambda}^{-1}, \\ (\underline{\Lambda}^{-\frac{1}{2}} \underline{A}) \sum_{i=1}^n (\underline{\Lambda}^{-\frac{1}{2}} \underline{A})^T = \underline{I}. \quad (27)$$

Substituting Eq. (26) into Eq. (6)

for $X \in \omega_2$:

* (3), p. 157. $M(j\omega)$ for the $N(\underline{\mu}, \sigma^2)$ case is calculated to illustrate the procedure used in general.

$$\begin{aligned}
U_2(Z) &= (Z+\tilde{v})^T \Lambda (Z+\tilde{v}) - Z^T Z + C \\
&= \sum_{i=1}^n \left\{ \lambda_i (z_i + \tilde{v}_i)^2 - z_i^2 - \ln \lambda_i \right\} \\
&\quad - 2 \ln [P(\omega_1)/P(\omega_2)] \\
&= \sum_{i=1}^n \left\{ (\lambda_i - 1) \left(z_i + \frac{\sqrt{\lambda_i} \mu_i}{\lambda_i - 1} \right)^2 - \left(\frac{\mu_i}{\lambda_i - 1} + \ln \lambda_i \right) \right\} \\
&\quad - 2 \ln [P(\omega_1)/P(\omega_2)] \quad (28)
\end{aligned}$$

Again, we have an expression in terms of n independent variables, z_i , each of which has zero mean and unit variance.

Proceeding just as in the U_1 case, we find that the characteristic function of U_2 has the same functional form as the one for U_1 . The only change we have to make in Eqs. (18), (19), (23), (24), and (25) is to substitute M_2 , F_{2i} , a_{2i} , b_{2i} , h_{2i} and Θ_{2i} for M_1 , F_{1i} , a_{1i} , b_{1i} , h_{1i} and Θ_{1i} . Instead of Eqs. (20), (21), and (22), a_{2i} , b_{2i} and h_{2i} are given by

$$a_{2i} = \lambda_i - 1, \quad (29)$$

$$b_{2i} = \frac{\sqrt{\lambda_i} \mu_i}{\lambda_i - 1}, \quad (30)$$

$$h_{2i} = \frac{-(a_{2i} b_{2i})^2}{1 + a_{2i}} + \ln \lambda_i. \quad (31)$$

IV. Inverse Transformation

Recognizing that the Fourier transform of a probability density is just the characteristic function with $-j$ replacing j , we can now proceed to find the distributions of U_1 and U_2 . In particular, we want to find the distributions, i.e. the integral of the density function, at U_1 and U_2 equal zero, since

$$\begin{aligned}
e_1 &= \text{Prob}(U_1 > 0) P(\omega_1) \text{ and} \\
e_2 &= \text{Prob}(U_2 < 0) P(\omega_2).
\end{aligned}$$

Denoting the probability density of U_k by $p_k(U_k)$, we can write the Fourier transform of this density as

$$M_k(-j\omega) = \int_{-\infty}^{\infty} p_k(U_k) \exp[-j\omega U_k] dU_k. \quad (32)$$

Taking the inverse,

$$\begin{aligned}
p_k(U_k) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} M_k(-j\omega) \exp[j\omega U_k] d\omega \\
&= \frac{1}{\pi} \int_0^{\infty} [\text{Re}[M_k(-j\omega)]] \cos \omega U_k - \text{Im}[M_k(-j\omega)] \sin \omega U_k d\omega, \quad (33)
\end{aligned}$$

where $M_k(-j\omega)$ is given by Eq. (18) or (23), replacing j by $-j$.

The recognition errors due to U_1 and U_2 can both be expressed in terms of the integral

$$D_k(U_k) = \int_{-\infty}^{\infty} p_k(U) dU, \quad k=1,2. \quad (34)$$

An application of basic transform theorems for integrals yields*

$$\begin{aligned}
D_k(U_k) &= \frac{M_k(0)}{2} + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{M_k(-j\omega)}{j\omega} \exp[j\omega U_k] d\omega \\
&= \frac{\text{Re}[M_k(0)]}{2} + \frac{1}{\pi} \int_0^{\infty} \left\{ \frac{\text{Im}[M_k(-j\omega)]}{\omega} \right. \\
&\quad \left. \cos \omega U_k + \frac{\text{Re}[M_k(-j\omega)]}{\omega} \sin \omega U_k \right\} d\omega.
\end{aligned} \quad (35)$$

Since we are only interested in $D_k(0)$, that is, the integral of $p_k(U_k)$ from $-\infty$ to 0 , Eq. (35) reduces to

$$\begin{aligned}
D_k(0) &= \frac{\text{Re}[M_k(0)]}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{\text{Im}[M_k(-j\omega)]}{\omega} d\omega \\
&= \frac{1}{2} + \frac{1}{\pi} \int_0^{\infty} \frac{i \sum_{i=1}^n |F_{ki}(\omega)|}{\omega} \sin \left(-\sum_{i=1}^n \Theta_{ki}(\omega) \right) d\omega. \quad (36)
\end{aligned}$$

From Eqs. (8) and (9), D_k 's are related to the recognition errors as follows:

* (4), p. 40.

$$e_1 = P(\omega_1) [1 - D_1(0)], \text{ and } e_2 = P(\omega_2) D_2(0). \quad (37)$$

Using numerical integration on Eq. (36), the recognition error can be found exactly in terms of the eigenvalues, λ_i , of Eq. (11), and the components, μ_i , of the transformed difference-of-means vector. As is seen in Eqs. (23) and (24), $|F_{ki}(j\omega)|$ is a monotonically decreasing function of ω , and, because of the multiplication of $n |F_{ki}(j\omega)|^2$, $\sum_{i=1}^n |F_{ki}(j\omega)|$ decreases quickly with ω , especially for large n . Also, ω appears in the denominator of the integrand, so the amplitude of the integrand decreases even more quickly with increasing ω . Therefore, the number of sampling points needed for convergence of the integral in numerical integration has been experienced to be relatively small. An example illustrating the application of these results will be given in the last section.

V. Approximations of the Density of U_k

The results of the previous section provide an exact solution for the recognition error, but require two numerical integrations by computer, so an approximation algorithm will be sought.

5.1 Gaussian Approximation

If $\Sigma_1 = \Sigma_2$, all λ_i 's are equal to one, and the density of U_k is Gaussian. Thus recognition error can be found exactly using normal density curves. When $\Sigma_1 \neq \Sigma_2$, but

all λ_i 's are close to one, the density of U_k may be approximated by a Gaussian curve having the mean and variance of U_k .

Although the Gaussian approximation gives us a simple algorithm with which to calculate recognition rate, it holds only when all λ_i 's are close to one. This comes mainly from the fact that the actual densities of U_1 and U_2 are essentially shifted causal functions with non-zero skewness, as will be discussed in the next paragraph. Since the higher order moments are also available from the characteristic function of U_k as represented by Eq. (18), it is theoretically possible to improve the approximation by using a series of Hermite polynomials whose coefficients are functions of the moments.^{*,**} However, our experience has indicated that the convergence of the series is very slow and that many higher moments are required to reach convergence, unless all λ_i 's are close to one. Since we expect, broad range of λ_i 's in general, a different density approximation will be proposed in the next paragraph.

5.2 Convolution of two modified gamma distributions:

Examining Eqs. (14) and (28), the density function of U_k should be obtained by convolving the densities of n non-central χ^2 variables having multiplicative constants, and

* (4), p. 233.

** (6), Fisher discusses advantages and disadvantages of this expansion.

adding a shift parameter.

Separating the U_k density into three parts,

$$V_{kr} = \sum_{a_{ki} \geq 0}^{n_{kr}} a_{ki} (\xi_{ki} + b_{ki})^2 \text{ for } a_{ki} \geq 0 \quad (38)$$

$$V_{ks} = \sum_{a_{kj} < 0}^{n_{ks}} a_{kj} (\xi_{kj} + b_{kj})^2 \text{ for } a_{kj} < 0 \quad (39)$$

$$\zeta = \sum_{i=1}^n \left(\mu_i \lambda_i + \frac{\mu_i^2}{\lambda_i - 1} \right) + 2 \frac{P(u_1)}{P(u_2)}, \quad (40)$$

($n = n_{kr} + n_{ks}$)

where a_{1i} , b_{1i} and ξ_{1i} are for U_1 , and given by Eqs. (20), (21) and $\xi_{1i} = y_i$; likewise, a_{2i} , b_{2i} and ξ_{2i} are for U_2 , and given by Eqs. (29), (30) and $\xi_{2i} = z_i$.

The density function of V_{kr} , $p_{kr}(U)$, is the convolution of n_{kr} densities of squared Gaussian variables having multiplicative constants. All n_{kr} densities lie above the positive U axis with $a_{ki} \geq 0$. Likewise, the density function of V_{ks} , $p_{ks}(U)$, is the convolution of n_{ks} densities of squared Gaussian variables with multiplicative constants. All n_{ks} densities lie on the negative U axis with $a_{kj} < 0$. Therefore, both $p_{kr}(U)$ and $p_{ks}(-U)$ ($p_{ks}(U)$ reflected to the positive side) could be reasonably approximated by a general gamma form, especially for large n_{kr} and n_{ks} , as follows:

$$g(U) = \begin{cases} \frac{U^\alpha e^{-U/\beta}}{\beta^{\alpha+1} \Gamma(\alpha+1)} & \text{for } U \geq 0 \\ 0 & \text{for } U < 0 \end{cases} \quad (41)$$

where Γ is a gamma function. The parameters α and β can be determined

so that the mean, η , and the variance, σ^2 , of the true distribution match those of the approximation. From the moment generating function of U_k , the true means and variances for $p_{kr}(U)$ and $p_{ks}(U)$ are obtained as the summation of n_{kr} terms and n_{ks} terms as follows:

$$\eta_{kr} = \sum_{a_{ki} > 0}^{n_{kr}} a_{ki} (1 + b_{ki}^2) \text{ for } p_{kr}(U) \quad (42)$$

$$\eta_{ks} = \sum_{a_{kj} < 0}^{n_{ks}} a_{kj} (1 + b_{kj}^2) \text{ for } p_{ks}(U) \quad (43)$$

$$\sigma_{kr}^2 = 2 \sum_{a_{ki} > 0}^{n_{kr}} a_{ki}^2 (1 + 2b_{ki}^2) \text{ for } p_{kr}(U) \quad (44)$$

$$\sigma_{ks}^2 = 2 \sum_{a_{kj} < 0}^{n_{ks}} a_{kj}^2 (1 + 2b_{kj}^2) \text{ for } p_{ks}(U). \quad (45)$$

Then, α and β are obtained from

$$\alpha + 1 = \frac{\eta^2}{\sigma^2} \text{ and } \beta = \frac{\sigma^2}{\eta}. \quad (46)$$

Thus, the density function of U_k , which is our final goal, is the convolution of two gamma densities with a constant shift: one is distributed on the positive side of the U axis, and the other is on the negative side.

Here, let us observe what the densities of U_1 and U_2 look like.

(1) The density of U_1 : Since

$a_{1i} = (1 - \frac{1}{\lambda_i})$, from Eq. (20), positive a_{1i} 's are all distributed between 0 and 1 (λ_i 's are always positive). On the other hand, negative a_{1j} 's could be large negative values with small λ_i 's. Therefore, unless all λ_i 's are close to one, $p_{1r}(U)$ is a relatively sharp density in comparison with $p_{1s}(U)$. Thus, the convolu-

* (4), p. 236.

tion of $p_{1r}(U)$ and $p_{1s}(U)$ becomes a skewed density. This was the reason why we had a poor estimation of recognition rate by assuming a Gaussian distribution for U_1 . When all λ_i 's are close to one, $p_{1r}(U)$ and $p_{1s}(U)$ become similar densities, and their convolution approaches a Gaussian density.

(2) The density of U_2 : For U_2 , $a_{21} = (\lambda_1 - 1)$ from Eq. (29). Therefore, $p_{2s}(U)$ becomes a relatively sharp density in comparison with $p_{2r}(U)$. Thus, again the convolution of $p_{2r}(U)$ and $p_{2s}(U)$ is skewed, and the skewed direction is opposite to the one of U_1 .

In the above discussion, it was postulated that the densities of U_k are approximated by the convolution of two gamma densities with a constant shift. However, the convolution is hard to obtain in an explicit mathematical expression, because α is not an integer in general. Since a calculating algorithm for recognition rate without approximation has already been established if numerical integration is allowed, it is useless to introduce numerical integration here for calculating the convolution of approximated distributions.

In order to get an explicit expression for the convolution of $p_{kr}(U)$ and $p_{ks}(U)$, we propose a modified gamma distribution as follows:

$$g'(U) = \begin{cases} \frac{(U-c)^{\gamma} e^{-\frac{U-c}{\delta}}}{\delta^{\gamma+1} \Gamma(\gamma+1)} & \text{for } U \geq c \\ 0 & \text{for } U < c \end{cases} \quad (47)$$

$\gamma = 0 \text{ or } 1.$

That is:

(1) Gamma density curves are roughly categorized into two types: one is $e^{-\frac{U}{\delta}}$, and the other is $U e^{-\frac{U}{\delta}}$, depending on whether α obtained by Eq. (46) is larger or smaller than a threshold value 0.35.

(2) In addition, the approximation function is shifted by c along the U axis, in order to match both its mean and variance to those of the true distribution.

The threshold value of α given here is a compromise value. It was chosen in an attempt to match the maximum value and location of the maximum value of the gamma density to the modified gamma approximation.

(1) Calculate η and σ of the true distribution by Eqs. (42) through (45).

(2) Calculate α by Eq. (46).

(3) $\gamma = 0$ if $\alpha \geq 0.35$, and $\gamma = 1$ if $\alpha < 0.35$.

(4) Calculate δ and c by the following equations.

$$\gamma+1 = \frac{(\eta-c)^2}{\sigma^2} \text{ and } \delta = \frac{\sigma^2}{\eta-c} \quad (48)$$

Eq. (48) is exactly the same as Eq. (46), except for the shift of the mean.

Then, the convolution of $p_{kr}(U)$ and $p_{ks}(U)$, $p_k^*(U)$, can be obtained as an explicit expression as follows:

$$\begin{aligned}
p_k^*(U) &= \int_{-\infty}^{+\infty} p_{kr}(\tau) p_{ks}(U-\tau) d\tau \\
&= \frac{1}{\delta_{kr}^{\gamma_{kr}+1} \delta_{ks}^{\gamma_{ks}+1} \Gamma(\gamma_{kr}+1) \Gamma(\gamma_{ks}+1)} \int_{\tau_0}^{+\infty} \\
&\quad \left\{ \frac{(\tau - c_{kr})^{\gamma_{kr}} (\tau - U - c_{ks})^{\gamma_{ks}}}{e^{\frac{-\tau - c_{kr}}{\delta_{kr}} - \frac{\tau - U - c_{ks}}{\delta_{ks}}}} d\tau \right\}, \quad (49)
\end{aligned}$$

where

$$\tau_0 = \begin{cases} c_{kr} & \text{for } t = U - (c_{kr} - c_{ks}) \leq 0 \\ U + c_{ks} & \text{for } t = U - (c_{kr} - c_{ks}) \geq 0. \end{cases} \quad (50)$$

Calculating Eq. (49) and letting $t = U - (c_{kr} - c_{ks})$, we find

$$\begin{aligned}
p_k^*(t) &= \begin{cases} \frac{\delta_{kr}^{\gamma_{kr}}}{(\delta_{kr} + \delta_{ks})^{\gamma_{kr}+1}} \\ \left[-\frac{t}{\delta_{ks}} + \frac{(\gamma_{kr} + \gamma_{ks}) \delta_{kr}}{\delta_{kr} + \delta_{ks}} \right]^{\gamma_{ks}} \end{cases} \quad (51) \\
&\quad e^{\frac{t}{\delta_{ks}}} \quad \text{for } t \leq 0; \\
&\quad \frac{\delta_{kr}^{\gamma_{kr}}}{(\delta_{kr} + \delta_{ks})^{\gamma_{ks}+1}} \\
&\quad \left[\frac{t}{\delta_{kr}} + \frac{(\gamma_{kr} + \gamma_{ks}) \delta_{ks}}{\delta_{kr} + \delta_{ks}} \right]^{\gamma_{kr}} \\
&\quad - \frac{t}{\delta_{kr}} \quad \text{for } t \geq 0.
\end{aligned}$$

Defining the distance d_k as

$$d_k = \zeta - (c_{kr} - c_{ks}), \quad (52)$$

we can find e_1 by integrating $p_1^*(t)$ from d_1 to $+\infty$, and e_2 by integrating

$p_2^*(t)$ from $-\infty$ to d_2 . The term d_k brings the shift parameter ζ back into the picture, and also accounts for the displacement of the U_k approximations by c_{kr} and c_{ks} . In general,

$$\begin{aligned}
D^*(d_k) &= \int_{-\infty}^{d_k} p_k^*(t) dt = \\
&= \begin{cases} \frac{\delta_{ks}^{\gamma_{ks}+1}}{(\delta_{kr} + \delta_{ks})^{\gamma_{kr}+1}} \\ \left[\frac{d_k + 1 + (\gamma_{kr} + \gamma_{ks}) \delta_{kr}}{\delta_{kr} + \delta_{ks}} \right]^{\gamma_{ks}} \end{cases} \\
&\quad \frac{d_k}{e^{\delta_{ks}}}, \quad d_k \leq 0; \\
&\quad 1 - \left(\frac{\delta_{kr}^{\gamma_{ks}+1}}{(\delta_{kr} + \delta_{ks})^{\gamma_{ks}+1}} \right) \\
&\quad \left[\frac{d_k + 1 + (\gamma_{kr} + \gamma_{ks}) \delta_{ks}}{\delta_{kr} + \delta_{ks}} \right]^{\gamma_{kr}} \\
&\quad \frac{d_k}{e^{\delta_{kr}}}, \quad d_k \geq 0,
\end{aligned} \quad (53)$$

where $D^*(d_k)$ is the approximation for $\text{Prob}(U_k \leq 0)$.

Thus the approximated values of recognition errors are

$$\begin{aligned}
e_1 &= P(w_1)[1 - D^*(d_1)] \quad \text{and} \\
e_2 &= P(w_2)[D^*(d_2)]. \quad (54)
\end{aligned}$$

VI. Example of Exact and Approximation Algorithms

The exact and approximate error calculations will be applied to data presented in Marill and Green's paper¹. This data arises from eight tests, or pattern vector components, used to distinguish between the handwritten letters A and B. The transformed pattern components, represented

by λ_i and μ_i , will be ordered according to two well-known separability criteria: the divergence, and Bhattacharyya's distance. Then 8 subsets of n best transformed pattern components will be formed for $n=1,2,\dots,8$, and the exact and approximate recognition rates calculated for each subset. The errors due to U_1 and U_2 in each case will be displayed. The results of ordering the transformed pattern components, represented by the λ_i 's and μ_i 's, from most important to least important are shown in Table I.

Best subsets of transformed pattern components are obtained by taking the first component in the ordering for the subset $n=1$, the first two components in the ordering for the subset $n=2$, etc., until all components are contained in the subset $n=8$. The values of divergence, Div., and Bhattacharyya's distance B , for these subsets are also shown in Table I, normalized with respect to the subset $n=8$. The exact recognition rate for each subset, obtained by numerical integration of Eq. (36), is shown in Table II, along with the recognition rates due to U_1 and U_2 .

Two approximations for U_k , one by Gaussian distributions and the other by modified gamma distributions, were obtained with the same subsets of pattern components, and recognition rate results are also shown in Table II for compar-

ison with the exact method. The results show that the modified gamma distributions give a very good approximation of the recognition rates (about 1%), although the normal distributions give poor results (about 9%).

Figure 1 shows the exact densities of $p_{1r}(U)$ and $p_{1s}(U)$ and their modified gamma approximations. The exact densities can be obtained by Eq. (36), where $M_1(j\omega)$ should be calculated by using $n_{1r} F_{1i}$'s and $n_{1s} F_{1i}$'s in Eq. (23) instead of using $n F_{1i}$'s. The same set of curves for the U_2 case is shown in Fig. 2. In Fig. 3, the actual and approximated densities of U_1 and U_2 for the 8 variable case are displayed together for comparison. The actual density was again obtained by taking the inverse Fourier transform of Eq. (36) using numerical integration, while the approximated densities came from Eq. (51).

From these curves and the comparison of recognition rates, the modified gamma distribution provides a simple and accurate method for approximating recognition rate. It also gives us a feeling for the role that individual pattern vector components, represented by λ 's and μ 's, play in the overall recognition rate.

VII. Summary

This paper has presented an algorithm with which to calculate the recognition rate exactly when

pattern vectors with Gaussian statistics are applied to an optimum Bayes' classifier. The statistics for each class may have different covariance matrices and a priori probabilities. The procedure is as follows:

(1) Recognition error is expressed in terms of the discriminant function U as a one-dimensional variable.

(2) With the aid of a transformation, the characteristic function of U is found as the product of one-dimensional characteristic functions.

(3) The distribution of U and, finally, the recognition error, can be obtained from the characteristic function by an inverse Fourier transformation.

The distributions of U are approximated by both the Gaussian distribution and a convolution of modified gamma distributions with a constant shift. The modified gamma distributions gave an explicit expression for recognition error in terms of the λ 's and μ 's,

and also gave good results when applied to the data from Marill and Green's example.

References

- (1) T. Marill and D.M. Green, "On the effectiveness of receptors in recognition systems," *IEEE Trans. on Information Theory*, Vol. IT-9, pp. 11-17, January 1963.
- (2) S. Kullback, Information Theory and Statistics, John Wiley and Sons, Inc., New York, p. 195, 1959.
- (3) S. Wilks, Mathematical Statistics, John Wiley and Sons, Inc., New York, 1963.
- (4) A. Papoulis, The Fourier Integral and its Applications, McGraw-Hill, New York, 1962.
- (5) T.W. Anderson, An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, Inc., New York, 1962.
- (6) A. Fisher, The Mathematical Theory of Probabilities, Vol. 1, The MacMillan Company, New York, pp. 188-234, 1923.

Divergence Orderings	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
	8.41	12.06	0.12	0.22	1.49	1.77	0.35	2.73
Bhattacharyya Distance Orderings	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8^*
	3.86	3.10	0.84	0.84	1.64	1.08	0.26	0.01

No. of components in subset (n)		1	2	3	4	5	6	7	8
Divergence of Subset	Actual	11.6	21.9	28.2	31.4	33.8	34.8	35.6	36.1
	% of Max.	32.1	60.7	78.1	87.0	93.6	96.4	98.7	100.0
B Distance of Subset	Actual	2.6	4.6	6.1	7.2	8.3	8.8	9.1	9.4
	% of Max.	27.7	48.9	64.9	76.6	88.3	93.6	96.8	100.0

* Since μ_1 always appears in the equations as μ_1^2 , only absolute values are tabulated.

Table I.

n		1	2	3	4 Div. 4B	5	6	7	8
Exact Recognition Rate	Prob ($U_1 \leq 0$)	0.941	0.972	0.973	0.974 0.978	0.980	0.983	0.984	0.985
	Prob ($U_2 \geq 0$)	0.782	0.907	0.932	0.950 0.957	0.969	0.974	0.976	0.978
	$R = \frac{1}{2} [\text{Prob } (U_1 \leq 0) + \text{Prob } (U_2 \geq 0)]^*$	0.861	0.940	0.953	0.962 0.968	0.975	0.978	0.980	0.981
Approximate Recognition Rate for Modified Gamma Density	Prob ($U_1 \leq 0$)	0.948	0.972	0.941	0.987 0.958	0.987	0.988	0.969	0.972
	Prob ($U_2 \geq 0$)	0.833	0.908	0.929	0.944 0.956	0.967	0.977	0.978	0.982
	$R = \frac{1}{2} [\text{Prob } (U_1 \leq 0) + \text{Prob } (U_2 \geq 0)]^*$	0.891	0.940	0.935	0.966 0.957	0.977	0.983	0.974	0.977
Approximate Recognition Rate for Gaussian Density, matching mean and variance	Prob ($U_1 \leq 0$)	0.975	0.995	0.828	0.865 0.855	0.884	0.891	0.898	0.901
	Prob ($U_2 \geq 0$)	0.793	0.857	0.866	0.874 0.881	0.888	0.894	0.896	0.899
	$R = \frac{1}{2} [\text{Prob } (U_1 \leq 0) + \text{Prob } (U_2 \geq 0)]^*$	0.884	0.926	0.847	0.870 0.868	0.866	0.893	0.897	0.900

$$^* P(\omega_1) = P(\omega_2) = \frac{1}{2}$$

Table II.

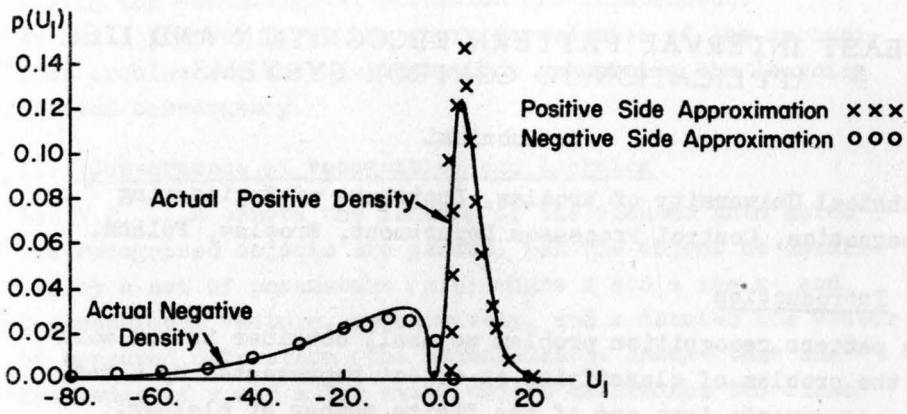


FIG. 1.

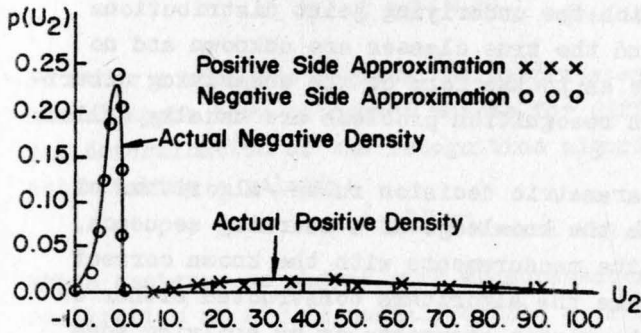


FIG. 2.

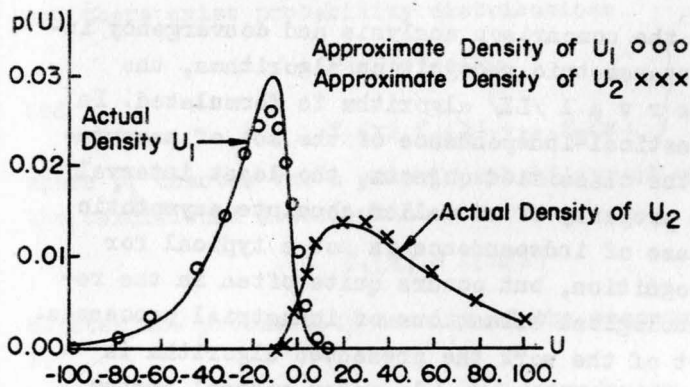


FIG. 3.

LEAST INTERVAL PATTERN RECOGNITION AND ITS APPLICATION TO CONTROL SYSTEMS

Z. Bubnicki

Technical University of Wrocław, Institute of Engineering Cybernetics, Control Processes Department, Wrocław, Poland.

I. Introduction

The pattern recognition problem we shall consider in the work is the problem of classifying an object represented by a set of measurements into one of the finite number of classes. In the problems of statistical pattern recognition there are many situations in which the underlying joint distributions of the measurements and the true classes are unknown and no assumption can be made as to the form of the underlying distributions. These pattern recognition problems are usually called nonparametric.

There exist many nonparametric decision rules /algorithms of recognition/ which use the knowledge of a learning sequence, i.e. the sequence of the measurements with the known correct classification. They are the algorithms constructed either on the basis of certain heuristic arguments or by applying some analytical methods¹⁻⁷/for example, the methods of stochastic approximation/.

In this work, via the comparison analysis and convergency investigation of nonparametric recognition algorithms, the least interval /LI/ algorithm is formulated. In the case of stochastic independence of the set of measurements describing the classified objects, the least interval algorithm has the property of so called absolute asymptotic optimality. The case of independence is not a typical for the character recognition, but occurs quite often in the recognition of technological situations of industrial processes. In the second part of the work the presented algorithm is applied in certain conception of a learning control system. It will be the automatic optimization of an industrial process when the measured parameters and disturbances characte-

rizing the technological situation are independent.

We shall present first the short formulation of the recognition problem and the considerations concerning the learning process convergency.

II. Convergence of recognition and learning

Let $1, 2, \dots, M$ denote the indexes of the classes into which the recognized objects are placed. Let the object be described by a set of parameters (x, a) where x and a are k - and p -dimensional vectors, respectively, and x denotes the vector of measured parameters (the measurement). Assume that the knowledge of x and a for every object determines the class to which the object belongs:

$$j = \Phi(x, a), \quad j = 1, 2, \dots, M \quad (1)$$

In other words, the regions $\{(x, a) : \Phi(x, a) = j\}$ in $(k+p)$ -dimensional space have no common points for different j .

The determination of the recognition algorithm consists in establishing a relation

$$i = \Psi(x), \quad i = 1, 2, \dots, M \quad (2)$$

which assigns an object with measurement x to class i . The realization of (2) is usually called a **r e c o g n i z e r**. We shall assume that the objects appear (are given) randomly from the populations corresponding to M specified classes and there exist probability distributions

$$\left. \begin{array}{l} p_1, p_2, \dots, p_M \\ \text{and} \\ f_1(x), f_2(x), \dots, f_M(x) \end{array} \right\} \quad (3)$$

where p_1 denotes the a priori probability of occurrence and the conditional distribution

$$f_1(x) = f(x/1)$$

denotes the probability density of the measurement x in class 1.

Knowing the distributions (3) and introducing some loss function $L(i, j)$, we may determine the optimal recognition algorithm (the optimum Bayes' decision rule)

$$i = \Psi^*(x) \quad (4)$$

minimizing the average risk

$$R[\Psi(\cdot)] \triangleq E_{X,j} [L(i=\Psi(x), j)] = \sum_{j=1}^M P_j \int_X L(\Psi(x), j) f_j(x) dx \quad (5)$$

where X is the measurement space. It is well known that for zero-unit loss function

$$L(i, j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

the optimum decision rule is the following. The object with measurements x is assigned to class i which results in the largest a posteriori probability

$$P_i(x) = \frac{P_i f_i(x)}{f(x)} = \frac{P_i f_i(x)}{\sum_{j=1}^M P_j f_j(x)}$$

In other words,

$$\left. \begin{aligned} &\Psi^*(x) = i \\ &\text{if } P_i f_i(x) > P_l f_l(x) \end{aligned} \right\} \quad (6)$$

, $l=1, 2, \dots, i-1, i+1, \dots, M$.

Then the average risk denotes the average probability of misrecognition which for the recognition algorithm (6) has the minimum value

$$R[\Psi^*(\cdot)] \triangleq R^*$$

Consider now the nonparametric recognition problem with a learning process. The distributions (3) are then unknown but the learning sequence is given:

$$(x_1, j_1), (x_2, j_2), \dots, (x_n, j_n) \triangleq \{s_n\} \quad (7)$$

where x_1 denote the independent measurements and j_1 the results of the correct classification according to (1). The realization of the algorithm (1) is usually called a teacher. The different methods of learning and recognition are reduced to the different methods of establishing the algorithm (2) on

the basis of the knowledge of (7), and consequently, to the different manners of constructing the relation

$$i = \Psi(\{s_n\}, x) \triangleq \Psi_n(x). \quad (8)$$

As a basis of comparison of the different algorithms (8) one may introduce the average probability of misrecognition

$$R_n \triangleq E_{\substack{\{s_n\}, \\ \{x_j\}}} \{L[\Psi(\{s_n\}, x), j]\} \quad (9)$$

or

$$R_n = E_{\{s_n\}} [R(\{s_n\})], \quad (10)$$

where

$$R(\{s_n\}) \triangleq \sum_{j=1}^M p_j \int_X L[\Psi_n(x), j] f_j(x) dx \quad (11)$$

and $L(i, j)$ denotes the zero-unit loss function.

Since the random vectors $x_1, x_2, \dots, x_n$ are independent, then, according to (10)

$$R_n = \sum \prod_{k=1}^n p_{j_k} \underbrace{\int_X \dots \int_X}_{n} R(\{s_n\}) \prod_{k=1}^n f_{j_k}(x_k) dx_k \quad (12)$$

where summing is distributed for all variations with repetitions $j_1, j_2, \dots, j_n$ constructing from the set of cardinal numbers $1, 2, \dots, M$ (the number of these variations is M^n).

Since the determination of R_n requires the knowledge of the distributions (3), the evaluation of some algorithm (8) and comparison with other algorithms may be done by assuming some concrete distributions (8).

The average probability of misrecognition may also be useful as a basis of the formulation and investigation of asymptotic properties of the algorithms (3). In the connection with it let us introduce the following definitions.

Definition 1

The algorithm (8) is *c o n v e r g e n t* if the sequence R_n is convergent. Define then

$$R = \lim_{n \rightarrow \infty} R_n.$$

Definition 2

The algorithm (8) is *a s y m p t o t i c a l l y o p t i m a l a.a.* for some distributions (3) if for these distribu-

tions $R = R^*$. Particularly, we may consider the asymptotic optimality of (8) in deterministic case. We shall call so a case in which the regions $\bar{D}_x^j = \{x : f_j(x) \neq 0\}$ have no common points for different j . Then

$$\Psi(x) = \Phi(x) = j \quad (13)$$

but the algorithm (13) is unknown.

Definition 3

The algorithm (8) will be called **a b s o l u t e l y a.a.** if it is a.a. for any distributions (3).

As an example, consider now two "heuristic" algorithms (8).

1. Nearest Neighbor (NN)

Then

$$\Psi_n(x) = j_p = i$$

if

$$\|x - x_p\| < \|x - x_t\|, \quad t=1, 2, \dots, p-1, p+1, \dots, n$$

for some norm $\|\cdot\|$. It was proved⁸ that the "nearest neighbor" rule is convergent and

$$R^* \leq R \leq R^* \left(2 - \frac{M}{M-1} R^*\right) \quad (14)$$

This algorithm is not absolutely a.a., but it is clear from (14) that the algorithm is a.a. in deterministic case ($R^*=0$) and if $p_1 = p_2 = \dots = p_M = \frac{1}{2}$, $f_1(x) = f_2(x) = \dots = f_M(x)$ for all x ($R^* = \frac{M-1}{M}$).

2. Nearest Mean (NM)

Then

$$\Psi_n(x) = i$$

if

$$\|x - \bar{x}_n^{(i)}\| < \|x - \bar{x}_n^{(l)}\|, \quad l=1, 2, \dots, i-1, i+1, \dots, M$$

where $\bar{x}_n^{(i)}$ denotes the arithmetical means of the measurements in the learning sequence, classified correctly into class i . Since the mean measurement $\bar{x}_n^{(i)}$ converges stochastically to the expected vectors $\bar{x}^{(i)}$, it is easy to show that

$$R = \sum_{j=1}^M p_j \sum_{\substack{i=1 \\ i \neq j}}^M \int_{D_x^{(i)}} f_j(x) dx$$

where $D_x^{(i)} = \{x : \|x - \bar{x}^{(i)}\| < \|x - \bar{x}^{(l)}\|, l \neq i\}$.

Hence, the MM algorithm is convergent but, since the region

$D_x^{(1)}$ may be not identical with the region $\bar{D}_x^1 = \{x: f_1(x) \neq 0\}$, the algorithm may be not a.a. in deterministic case (and consequently, it is not absolutely a.a.).

In the next section we shall formulate a new algorithm (8), absolutely a.a. when the particular components of the random vector x are stochastically independent.

III. Least Interval Algorithm (II)

The presented approach consists in the construction of Bayes-like optimum decision rule (6) in which the empirical probability and probability density are used instead of the p_1 and $f_1(x)$, respectively. These empirical distributions are evaluated on the basis of the knowledge of the learning sequence. The convergence of the empirical distributions to p_1 and $f_1(x)$ implies the absolute asymptotic optimality of the presented algorithm.

We shall present now some manner of establishing the empirical density. Consider first the onedimensional case. Let x be a continuous random variable with the probability density $f(x)$ and $x_1, x_2, \dots, x_n$ denote the observed values of x with the increasing ordering. We shall use the following empirical densities:

$$1. \quad (1) \quad f_n(x) = \begin{cases} 0 & x \leq x_0 \\ \frac{1}{n(x_{r+1} - x_r)} & x_r < x \leq x_{r+1}, \quad r=0, 1, \dots, n \\ 0 & x > x_n \end{cases} \quad (15)$$

where $x_0 = 2x_1 - x_2$.

$$2. \quad (2) \quad f_n(x) = \begin{cases} 0 & x \leq x_1 \\ \frac{1}{n(x_{r+1} - x_r)} & x_r < x \leq x_{r+1}, \quad r=1, 2, \dots, n+1 \\ 0 & x > x_{n+1} \end{cases} \quad (16)$$

where $x_{n+1} = 2x_n - x_{n-1}$.

$$3. \quad f_n^{(3)}(x) = \begin{cases} 0 & x \leq x_1 \\ \frac{1}{(n-1)(x_{r+1} - x_r)} & x_r < x \leq x_{r+1}, \quad r=1, 2, \dots, n \\ 0 & x > x_n \end{cases} \quad (17)$$

The respective three forms of the empirical distribution function are shown in Fig.1. In this figure the empirical distribution

$$F_n(x) = \begin{cases} 0 & x \leq x_1 \\ \frac{k}{n} & x_r < x \leq x_{r+1}, \quad r=1, 2, \dots, n \\ 0 & x > x_n \end{cases} \quad (18)$$

is shown also.

Consider now the recognition problem in the onedimensional case. Let

$$x_{n,1}^{(1)}, x_{n,2}^{(1)}, x_{n,3}^{(1)}, \dots, x_{n,n_1}^{(1)} \quad (19)$$

denote the increasing ordering of the measurements from the learning sequence which belong to class 1 and let

$$\Delta_n^{(1)}(x) = x_{n,r+1}^{(1)} - x_{n,r}^{(1)}$$

denote the length of the interval in which the measurement x lies. Therefore,

$$x_{n,r}^{(1)} < x \leq x_{n,r+1}^{(1)}.$$

Then, substituting the empirical probabilities $p_{1,n}$ and empirical densities $f_{1,n}(x)$ instead of p_1 and $f_1(x)$, respectively, we obtain

$$p_{1,n} f_{1,n}^{(1,2)}(x) = \frac{n_1}{n} \frac{1}{n_1 \Delta_n^{(1)}(x)},$$

$$p_{1,n} f_{1,n}^{(3)}(x) = \frac{n_1}{n} \frac{1}{(n_1-1) \Delta_n^{(1)}(x)}.$$

The algorithm (6) is now reduced to the following rules.

$$1.2. \quad \Psi_n(x) = 1 \quad (20)$$

if

$$\Delta_n^{(1)}(x) < \Delta_n^{(l)}(x), \quad l=1,2,\dots,i-1,i+1,\dots,M$$

In the case 1. the sequence (19) is expanded by $x_{n,0}^{(1)}$ and in the case 2. by $x_{n,n+1}^{(1)}$.

$$3. \quad \Psi_n(x) = 1 \quad (21)$$

if

$$x_n^{(1)} \cdot \Delta_n^{(1)}(x) < x_n^{(l)} \cdot \Delta_n^{(l)}(x), \quad l=1,2,\dots,i-1,i+1,\dots,M$$

where

$$x_n^{(1)} = \frac{n_1^{-1}}{n_1}.$$

The rule is then the following. For each class we determine the length of the interval in which x lies and then we classify x into the class for which this interval is the least. Comparing the intervals in the case 3. we multiply the lengths by the weight coefficients $x_n^{(1)}$, depending on the number of measurements from particular class in the learning sequence. The rules 1,2,3 may be considered as three modifications of the recognition algorithm which we shall call *least interval algorithm* (LI). If for some class l the measurement x lies in an infinite interval then $\Delta^{(l)}(x) = \infty$. If $\Delta^{(l)}(x) = \infty$ for each class then LI algorithm does not give the decision and may be replaced by some other algorithm (for example, NN). The probability of such a situation converges to zero for $n \rightarrow \infty$.

For $M=2$ (the rule 1), the LI recognition procedure is illustrated in Fig.2. Since the interval in which the point x lies in class 2 (the axis b) is less than in class 1 (the axis a), the point x is classified into class 2.

The short proof of the absolute asymptotic optimality of the LI algorithm is presented in Appendix.

Consider now the multidimensional case and assume that the particular components of the random vector x are independent.

Then

$$f_1(x) = f_1^{(1)}(x^1) f_1^{(2)}(x^2) \dots f_1^{(k)}(x^k), \quad l=1,2,\dots,M$$

where $f_1^{(s)}(x^s)$, $s=1,2,\dots,k$ denotes the onedimensional probability density for sth component and

$$f_{1,n}(x) = \prod_{s=1}^k f_1^{(s)}(x^s)$$

The onedimensional densities $f_1^{(s)}(x^s)$ are established in the same way as previously. We obtain then

$$p_{1,n} \cdot f_{1,n}^{(1,2)}(x) = \frac{n_1}{n} \frac{1}{n_1^k \Delta_n^{(1)} V(x)},$$

$$p_{1,n} \cdot f_{1,n}^{(3)}(x) = \frac{n_1}{n} \frac{1}{(n_1-1)^k \Delta_n^{(1)} V(x)}$$

where

$$\Delta_n^{(1)} V(x) = \prod_{s=1}^k \Delta_n^{(1)}(x^s)$$

denotes the volume of the multidimensional interval in which the point x lies. Finally, we obtain the following II algorithm

$$\Psi_n(x) = i$$

if

$$x_n^{(i)} \Delta_n^{(i)} V(x) < x_n^{(1)} \Delta_n^{(1)} V(x), \quad i=1,2,\dots,i-1,i+1,\dots, M \quad (22)$$

...,M where

in the case 1 and 2

$$x_n^{(1)} = n_1^{k-1},$$

in the case 3

$$x_n^{(1)} = \frac{(n_1-1)^k}{n_1}.$$

The II procedure is illustrated in Fig. 3 (the rule 1, $M=2$, $k=2$). Since the interval in class 1 is less than in class 2 and $n_1=n_2$, the point x is classified into class 1.

The simplified block diagram of the II algorithm (the structure of the recognizer) is shown in Fig. 4.

IV. Application in learning control system

We shall present now some conception of a static optimization system in which the II algorithm will be used for the recognition of technological situations. Consider an industrial

process described by

$$v = F(u, x, a) \quad (23)$$

where x is the vector of measured parameters, a is the vector of unmeasured parameters, u is the control γ -vector and v is a performance index which is to be minimized. Assume that for all x F is a convex function and there exists the vector

$$u^* = \varphi(x, a) \quad (24)$$

which minimizes v . Assume also that the controlling system, receiving the measurement x , determines and gives on the process input one of the M possible control vectors $u^{(1)}, u^{(2)}, \dots, u^{(M)}$. Let $u^{(j)}$ be the nearest vector to u^* in the sense of some norm, i.e.

$$\|u^* - u^{(j)}\| < \|u^* - u^{(l)}\|, \quad l=1, 2, \dots, j-1, j+1, \dots, M.$$

Thus, for each pair of the vectors (x, a) there exists the respective control vector $u^{(j)}$, defined by the index j :

$$\Phi(x, a) = j.$$

The pair (x, a) may be called a technological situation of the process, and the action of the controlling system, assigning the index i (and consequently $u^{(i)}$) to the measurement x , may be considered as the recognition of (x, a) knowing the measurement x . To design such controlling system (the recognizer) it is necessary to determine its recognition algorithm $i = \Psi(x)$.

Assume that x and a may change and denote by x_n and a_n their values in the recognition moments. Suppose that x_n and a_n change randomly and are independent for different n . The probability distributions (independent upon n) are unknown. It means that the a priori distributions (3) are unknown. To obtain the learning sequence one may apply the optimizer which for the particular (x_n, a_n) reduced the process input to the respective u_n (if the technological situations change slowly and it is assumed they are constant in the time of optimization) and then the respective index j_n is determined. The block diagram of the system is shown in Fig.5. After ending the learning process the switch changes the position and the process is controlled via the recognizer. For the complete

designing of the system it is necessary to determine the algorithms for the particular blocks.

I. Optimizer

Establishing the optimizer algorithm one must assure the convergence of the optimization process. Assume that the optimization algorithm is the following

$$u_{m+1} = u_m - K \cdot y_m \quad (25)$$

where K is $\nu \times \nu$ matrix with constant coefficients, u_m is the input vector in the m th moment of optimization, the s th component of y_m is the following

$$y_m(s) = \frac{F(u_m + \alpha_s, x, a) - F(u_m - \alpha_s, x, a)}{2\delta}, \quad s=1, 2, \dots, \nu$$

and α_s denotes the vector with zero components except the s th component equal to δ (the value of a trial step). Denote

$$y_m = H(u_m, x, a) = G(u_m, x, a)u_m.$$

Then the equation (25) takes the form

$$u_{m+1} = u_m [I - KG(u_m, x, a)] \quad (26)$$

and the convergency condition is the following⁹⁻¹². If for some norm $\|\cdot\|$ and some $0 < \varepsilon < 1$

$$\|I - KG(u, x, a)\| < \varepsilon \quad (27)$$

for all u , then for every u_0 $u_m \rightarrow u^*$.

Knowing x and the region of the variation of a , we may determine for some ε (arbitrary near to 1) the values of K satisfying (27) (the block K in Fig.5). It may be shown that, for symmetric matrix $I - KG(u)$ the condition (27) may be reduced to the form¹¹

$$|\lambda_s(u)| < \varepsilon, \quad s=1, 2, \dots, \nu \quad (28)$$

where $\lambda_s(u)$ are the eigenvalues of $I - KG(u)$.

II. Recognizer

If the measured parameters (the components of x) are independent, the recognizer may realize the II algorithm, assuring in this way the asymptotic optimality of learning process. For the measurement x , knowing the learning sequence, the recognizer computes the volumes ΔV of the intervals corresponding to the control vectors $u^{(1)}, u^{(2)}, \dots, u^{(M)}$ and chooses

$u^{(i)}$ such that

$$x_n^{(i)} \Delta_n^{(i)} V(x) < x_n^{(1)} \Delta_n^{(1)} V(x), \quad l=1, 2, \dots, i-1, i+1, \dots, M \text{ (see 22).}$$

III. Block B

The block B establishes the indexes j_n for the technological situations (x_n, a_n) on the basis of the result of optimization. The algorithm of the block B is the following

$$u_n^* \Rightarrow j_n$$

if

$$\|u_n^* - u^{(j_n)}\| < \|u_n^* - u^{(1)}\|, \quad l=1, 2, \dots, j_n-1, j_n+1, \dots, M.$$

...

V. Conclusions

On the basis of the investigation of convergency and optimality of the learning process in nonparametric pattern recognition problem, the least interval recognition algorithm has been formulated in this work. This algorithm is absolutely asymptotically optimal if the particular measured parameters are independent. Presented approach, based on the constructing of empirical probability distributions, may be extended for the case in which the components of the measured vector are dependent but the obtained algorithm will be much more complicated.

II algorithm has been applied in some conception of a learning control system described in the work. The algorithms of particular blocks in the learning system may be realized by separate devices or by one controlling digital computer.

Appendix

Let us denote

$$D_n = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|$$

where $F_n(x)$ is defined by (18). According to the Glivenko's theorem, D_n converges to zero with probability 1 (w.p.1):

Since

$$|F_n^{(1,2,3)}(x) - F(x)| \leq D_n + \frac{1}{n},$$

$f_n^{(1,2,3)}(x)$ converges to $F(x)$ uniformly w.p.1. According to the theorem of Levy and Cramer, the respective sequence of the characteristic functions $\varphi_n^{(1,2,3)}(t)$ converges uniformly w.p.1 to $\varphi(t)$ corresponding to $F(x)$. Hence,

$$f_n^{(1,2,3)}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jtx} \varphi_n^{(1,2,3)}(t) dt \xrightarrow{w.p.1} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-jtx} \varphi(t) dt = f(x).$$

It is easy to show that this convergence is uniform. Since $f_{n,1}(x)$ converge uniformly w.p.1 to $f_1(x)$ and $p_{1,n}$ converge to p_1 w.p.1, the LI algorithm $\Psi_n^*(x)$ converges uniformly to $\Psi^*(x)$ defined by (6). Define

$$R_n(x) \triangleq E_{\{s_n\},j} [L(\Psi_n(x), j)].$$

Then

$$\lim_{n \rightarrow \infty} R_n(x) = R^*(x),$$

where
$$R^*(x) = E_{\{s\},j} [L(\Psi^*(x), j)]$$

and the convergence is uniform. Hence,

$$\lim_{n \rightarrow \infty} R_n = \lim_{n \rightarrow \infty} \int_X R_n(x) f(x) dx = \int_X R^*(x) f(x) dx = R^*.$$

References

1. G. Sebestyen, "Decision Making Processes in Pattern Recognition". New York, Macmillan, 1962.
2. N. Abramson and D. Braverman, "Learning to recognize patterns in a random environment", IRE Trans on Information Theory, vol.8, pp. 58-62, September 1962.
3. N. Nilsson, "Learning Machines". New York, McGraw-Hill, 1965.
4. M.A. Aizerman and E.M. Braverman, L.I. Rozonoer, "Teoreticheskiye osnovy metoda potentsialnykh funktsiy..." Awtomatika i Telemekhanika No 6, 1964 (in Russian).
5. J.Z. Cypkin, "Adaptatsiya, obucheniye i samoobucheniye w awtomaticheskikh sistemach", Awtomatika i Telemekhanika No. 1, pp. 23-61, 1966 (in Russian).

6. J. Seidler, "Optimierung der Information Verarbeitung", Verlag Technik, 1967.
7. S. Wegrzyn, "Sterowanie procesami technologicznymi w oparciu o rozpoznawanie ich stanów." IV National Automation Conference, vol. 1, 1967 (in Polish).
8. T.M. Cover and P.E. Hart, "Nearest neighbor pattern class classification", IEEE Trans. on Information Theory, vol.13, pp. 21-27, January 1967.
9. Z. Bubnicki, "On the stability condition of nonlinear sampled-data systems", IEEE Trans. on Automatic Control, vol.9, July 1964.
10. Z. Bubnicki, "Ob usloviach schodimosti awtomatičeskowo poiska w diskretnych sistemach ekstremalnowo regulirivanija", Awtomatika i Telemekhanika No. 10, 1967 (Automation and Remote Control) (in Russian).
11. Z. Bubnicki, "On the linear conjecture in deterministic and stochastic stability of discrete systems", IEEE Trans. on Automatic Control, vol. 13, April, 1968.
12. Z. Bubnicki, "Zbieżność procesów automatycznej aproksymacji w układach dyskretnych", Zeszyty Nauk. Polś Wroc., Automatyka No 5, Wrocław, 1966 (in Polish).
13. M. Fisż, "Rachunek prawdopodobieństwa i Statystyka Matematyczna", Warszawa, PWN, 1954 (in Polish).

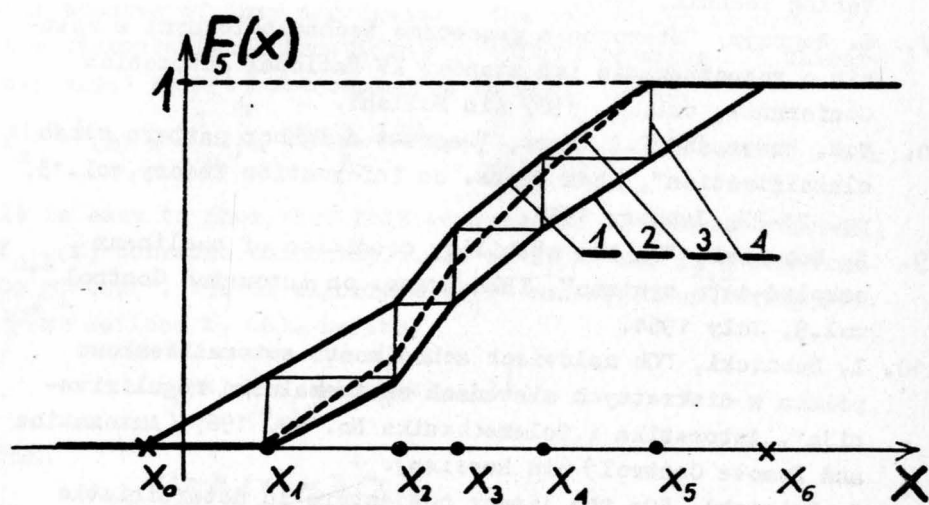


Fig.1. Empirical distribution functions: $F_n^{(1,2,3)}$ and $F_n(4)$.

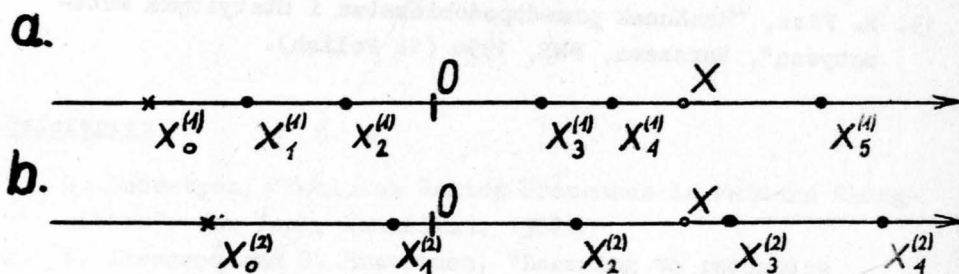


Fig.2. Illustration of LI algorithm for $k=1$.

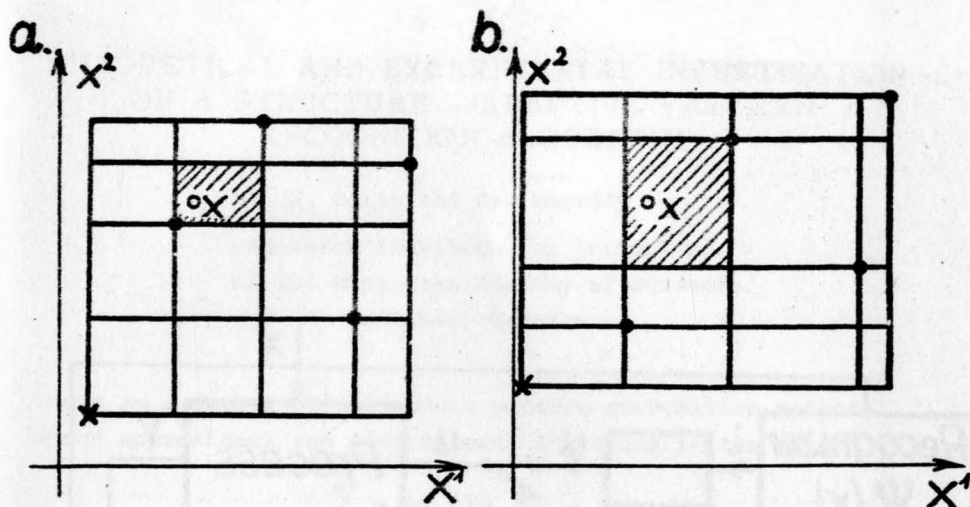


Fig.3. Illustration of LI algorithm for $k=2$, a - class 1, b - class 2.

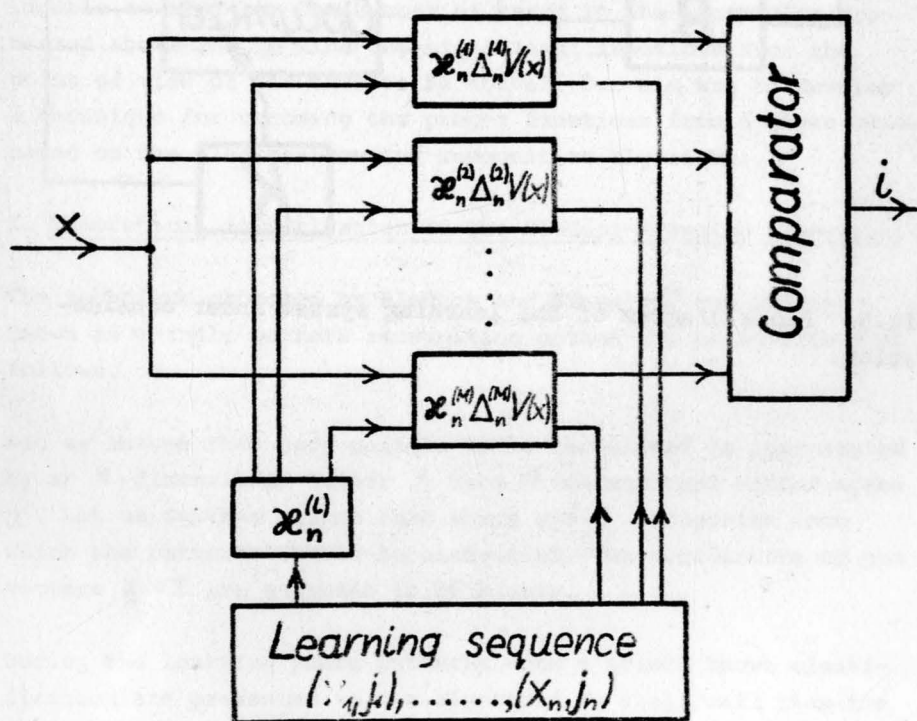


Fig.4. Simplified structure of the recognizer.

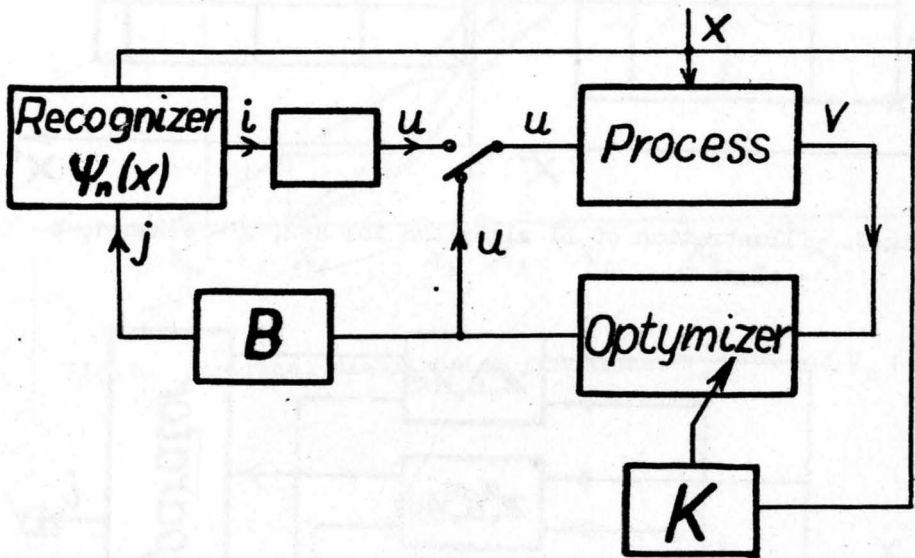


Fig.5. Block diagram of the learning system under consideration.

THEORETICAL AND EXPERIMENTAL INVESTIGATION OF A STRUCTURE - ADAPTIVE PATTERN RECOGNITION ALGORITHM

J. Barát and Gy. Muszély

Research Institute for Automation
of the Hungarian Academy of Sciences
Budapest, Hungary

There are a number of well-known pattern recognition methods which approximate the discriminant functions in the form:

$$\sum_{i=1}^m c \cdot \varphi_i(X)$$

/1/

where the functions $\varphi_i(X)$ are given and determine the "structure" of the pattern recognizer. But there is a profound interest in being able to minimize the number of terms in the expression presented above and to find the significant functions from the point of view of the task to be solved. Our aim was to develop a technique for choosing the proper functions from a given class based on the Bledsoe-Browning recognition algorithm.

1. Theoretical investigation of the Bledsoe-Browning algorithm.

The technique proposed by Bledsoe and Browning¹ and widely known as n -tuple pattern recognition method can be described as follows.

Let us assume that each pattern to be recognised is represented by an N -dimensional vector $\underline{X}$ from N -dimensional vector space $\mathcal{X}$. Let us further assume that there are n categories into which the patterns should be classified. The coordinates of the vectors $\underline{X} \in \mathcal{X}$ are supposed to be binary.

During the learning phase patterns with a priori known classification are presented to the algorithm. We shall call them the training sequence which is drawn from a finite sample

$S_i (i = 1, 2, \dots, s)$ of each of the categories.

Let us form p -tuples from the N components of $\underline{X}$. Let K be the number of p -tuples. Each p -tuple has 2^p binary values corresponding to the 2^p possible binary patterns.

Forming p -tuples and observing the binary values of this group of coordinates can be interpreted as a transformation performed by forming all the possible minterms from the elements of the group regarding them as boolean variables.

Thus, we can write

$$F(\underline{X}) = \{f_1(\underline{X}), f_2(\underline{X}), \dots, f_K(\underline{X})\} = \\ = \{\varphi_{1,0}(\underline{X}), \dots, \varphi_{1,2^p-1}(\underline{X}), \varphi_{2,0}(\underline{X}), \dots, \varphi_{2,2^p-1}(\underline{X}), \dots, \varphi_{K,2^p-1}(\underline{X})\} \quad /2/$$

where

$$\varphi_{i,j}(\underline{X}) = [x_{i,p}]^{\alpha_p} \& [x_{i,p-1}]^{\alpha_{p-1}} \& \dots \& [x_{i,1}]^{\alpha_1} \quad /3/$$

where

$$[x]^\alpha = \begin{cases} x & \text{if } \alpha = 1 \\ \bar{x} & \text{if } \alpha = 0 \end{cases}$$

and

$$\alpha_p \alpha_{p-1} \dots \alpha_1 \sim j; \quad j = 0, 1, \dots, 2^p - 1$$

are the different binary values from 0 to $2^p - 1$.

Let us consider for sake of simplicity the case, when the number of categories is two $/n = 2/$. Assume we have the sample sets $\{S_A, S_B\}$ from each of the categories A and B .

Then the Bledsoe-Browning method of learning consists of simple accumulation of the transformed by $F(\underline{X})$ data received from the learning sequence.

Denoting $F(\underline{X})$ by $\underline{Y}$ we can write the recursive expression:

$$\underline{W}_0 = \underline{0} \quad / \underline{0} \text{ denotes the zero vector} /$$

$$\underline{W}_n = \underline{W}_{n-1} + \underline{Y} \quad \text{if } \underline{X} \in A$$

$$\underline{W}_n = \underline{W}_{n-1} - \underline{Y} \quad \text{if } \underline{X} \in B$$

/4/

When the learning phase is completed the decision rule for an unknown pattern $\underline{X}$ is

$$g(\underline{X}) = (\underline{W}^*, \underline{F}(\underline{X})) \geq 0 \quad /5/$$

where $(\underline{X}, \underline{Y})$ denotes the inner product of vectors $\underline{X}$ and $\underline{Y}$.

Clearly in this case the decision surface is a hyperplane going through the origin and orthogonal to the straight line connecting the weight vectors of the two categories A and B.

The learning process (3) is a procedure without correction and thus there can not be proved anything like the Perceptron's convergence theorem². But if we consider the normalized components of the weight vector $\underline{W}^*$:

$$\frac{1}{s} w_{ij}^* = \frac{1}{s} \left[\sum_{\underline{X} \in A} \varphi_{ij}(\underline{X}) - \sum_{\underline{X} \in B} \varphi_{ij}(\underline{X}) \right] \quad i=1,2,\dots,s \quad /6/$$

where s is the sample size of each S_A and S_B , and supposing the $\varphi_{ij}(\underline{X})$ to be independent random variables drawn from the same distribution then by generalization of Bernoulli's theorem the sample mean converges in stochastic sense to the expectation of $\varphi_{ij}(\underline{X})$:

$$\frac{1}{s} \sum_{\underline{X} \in A} \varphi_{ij}(\underline{X}) \Rightarrow E(\varphi_{ij}(\underline{X})) \quad /7/$$

It is easy to show, that once the mean value and variance exists for $\underline{X}$ the same is true for $\varphi_{ij}(\underline{X})$. These arguments and the independence assumption leads us to the assertion, that by increasing the sample size $s \rightarrow \infty$ the process converges in stochastic sense to the solution vector if the distributions of the categories are symmetric. Here the "solution vector" means the optimal discriminating hyperplane orthogonal to the straight line connecting the mean vectors and going through the origin in the sense of minimal error rate.

As it has been proved³ the class of solvable problems for this special kind of ϕ -functions, as in case of the Perceptron machine⁴, is closed under permutation of the coordinates of the vectors $\underline{X} \in \mathcal{I}$.

This remarkable property of the boolean functions used in the Bledsoe-Browning algorithm makes the method to utilize only the number of common elements /0's and 1's/ among the learning patterns and the pattern to be recognized. Therefore, in general, one can not hope to extract by this method any metric or topological properties of the patterns unless these properties occur in the sample by "covering" the same points of the receptor space /the same coordinates of $\underline{X}$ /. So the properties "noticable" by this method must be invariant to permutations of the coordinates of the vector space $\mathbf{I}$.

Strictly speaking the expectation of a single term in the discriminant function (4), if the elements of the p -tuples are chosen by equal probability densities⁴

$$E[w_{ij}^* \cdot \varphi_{ij}(\underline{Y})] = E\left[\sum_{\underline{X} \in A} \varphi_{ij}(\underline{X}) \cdot \varphi_{ij}(\underline{Y}) - \sum_{\underline{X} \in B} \varphi_{ij}(\underline{X}) \cdot \varphi_{ij}(\underline{Y})\right] = \quad /8/$$

$$= \left[\frac{\sum_{\underline{X} \in A} \binom{m(\underline{X}, \underline{Y})}{p} - \sum_{\underline{X} \in B} \binom{m(\underline{X}, \underline{Y})}{p}}{\binom{N}{p}} \right],$$

where N - the dimensionality of the space
 p - the length of the p -tuple
 $m(\underline{X}, \underline{Y})$ - number of common elements for $\underline{X}$ and $\underline{Y}$
 /both 0's and 1's/

Although this fact might be rather annoying in cases of practical importance the class of functions used here is quite fortunate from point of view of the number of homogeneously linear dichotomies⁵ instrumentable by the help of them.

Using the well-known result of T.Cover⁵ an upper bound of the number of homogeneously linear dichotomies achievable by our ϕ -function can be found:

$$\phi(N, M) = L(N, 2^n) = \begin{cases} 2 \sum_{i=0}^{2^n} \binom{N-1}{i} & \text{for } N > 2^n \\ 2^N & \text{for } N \leq 2^n \end{cases} \quad /9/$$

This equation is correct only if every k -element subset of transformed by (2) vectors are linearly independent for all $k \leq 2^n$. In other cases it provides a useful upper bound for the number in question.

Following T. Cover's suggestion we can calculate the capacity of our ϕ -machine

$$C = 2(M+1) = 2(2^n + 1) \quad /10/$$

which in comparison with general quadric surfaces

$(C = (n+1)(n+2))$ gives a sensible improvement even for relatively small n 's.

2. Theoretical investigation of the structure adaptation technique.

Let us consider a given /finite or any infinite/ class of functions

$$\left\{ \varphi_i(\underline{X}) \right\}_{i \in I} \quad /11/$$

where $\underline{X}$ is an element of the above defined measurement /receptor/ space $\mathcal{I}$ and I is some index set. We would like to choose m functions from this class for our discriminant function (1).

Let define a measure of efficiency of a given function from (11) $h(\varphi_i)$ which shows how useful this function for our task is. For this purpose it is natural to consider the probability of the correct decision when it was made only on the basis of this function.

Let assume there exist the following probabilities:

$$\left(\begin{array}{ll} P[\varphi_i(\underline{X}) > 0, \underline{X} \in A] = p_i & P[\varphi_i(\underline{X}) > 0, \underline{X} \in B] = q_i \\ P[\varphi_i(\underline{X}) \leq 0, \underline{X} \in A] = p'_i & P[\varphi_i(\underline{X}) \leq 0, \underline{X} \in B] = q'_i \end{array} \right) \quad /12/$$

where clearly $p_i + p'_i + q_i + q'_i = 1$.

If we wanted to classify on the basis of the sign of $\varphi_i(X)$ then the strategy which provides the minimal error probability would be in case of $\varphi_i(X) > 0$ to classify X into A or B, or in case $\varphi_i(X) \leq 0$ into B or A according to which of $p_i + q_i'$ and $q_i + p_i'$ is greater. /In case of $p_i + q_i' = q_i + p_i'$ we can decide arbitrary./ We have the probability of correct decision:

$$h(\varphi_i) = \max(p_i + q_i', q_i + p_i') \quad /13/$$

where $h(\varphi_i)$ is in the interval $\left[\frac{1}{2}, 1\right]$.

This probabilities are unknown and all we can do to estimate them from the sample. Suppose we have sample consisting of k elements then we can form according to the matrix (12) the relative frequencies:

$$\begin{pmatrix} \frac{a_i}{k} & \frac{b_i}{k} \\ \frac{a_i'}{k} & \frac{b_i'}{k} \end{pmatrix} \quad /14/$$

It can be proved that on the basis of the strong law of large numbers the expression

$$\hat{h}(\varphi_i) = \frac{1}{k} \max(a_i + b_i', b_i + a_i') \quad /15/$$

converges with probability 1 to $h(\varphi_i)$ when $k \rightarrow \infty$.

The selection of efficient functions by mean of the above defined measure can be performed as follows.

Let be given a threshold value

$$\frac{1}{2} \leq c < 1.$$

Then choosing from the class $\{\varphi_i\}_{i \in I}$ functions at random and calculating the $\hat{h}(\varphi)$ value we keep only those for which $\hat{h}(\varphi_i) > c$. After finding m function this manner the search is completed. Now we can begin the process of estimating the weight vector $\underline{W}$ usually considered as the learning phase of the re-

cognition algorithm.

We state the following theorem: Providing there exists a probability distribution on the index set I of $\{\psi_i\}_{i \in I}$ and we draw a function ψ_i according to this distribution and for this function $\hat{h}(\psi_i) > c$ and we perform the classification only by this single function then the probability of correct decision is a monotone function of c .

Proof:

Let be $h(\psi_i)$ random variable given on the index set I , with distribution function:

$$F(t) = P(h(\psi_i) < t) \quad \frac{1}{2} \leq t \leq 1 \quad /16/$$

$$\text{If } c \leq \frac{1}{2} \quad \text{then } P(\hat{h}(\psi_i) \geq c) = 1. \quad /17/$$

$$\begin{aligned} \text{If } c > \frac{1}{2} \quad \text{then } P(\hat{h}(\psi_i) \geq c) &= P\left(\frac{1}{k}(a_i + b_i) \geq c\right) = \\ &= \int_{\frac{1}{2}}^1 \sum_{\frac{a}{k} \geq c} \binom{k}{a} \left[t^a (1-t)^{k-a} + t^{k-a} (1-t)^a \right] dF(t) \end{aligned} \quad /18/$$

For a given ψ_i the probability of the correct decision H for the condition mentioned above

$$\begin{aligned} P(H/\psi_i, \hat{h}(\psi_i) \geq c) &= \frac{P(H, \hat{h}(\psi_i) \geq c/\psi_i)}{P(\hat{h}(\psi_i) \geq c)} = \\ &= \frac{P(H/\psi_i) \cdot P(\hat{h}(\psi_i) \geq c/\psi_i)}{P(\hat{h}(\psi_i) \geq c)} = \frac{h(\psi_i) \cdot P(\hat{h}(\psi_i) \geq c/\psi_i)}{P(\hat{h}(\psi_i) \geq c)} \end{aligned} \quad /19/$$

The probability of correct decision when selecting the function ψ_i from our class $\{\psi_i\}_{i \in I}$ at random:

$$\begin{aligned} P(H/\hat{h}(\psi_i) \geq c) &= \frac{\int_{\frac{1}{2}}^1 t P(\hat{h}(\psi_i) \geq c/h(\psi_i) = t) dF(t)}{P(\hat{h}(\psi_i) \geq c)} = \\ &= \frac{\int_{\frac{1}{2}}^1 t \sum_{\frac{a}{k} \geq c} \binom{k}{a} \left[t^a (1-t)^{k-a} + t^{k-a} (1-t)^a \right] dF(t)}{\int_{\frac{1}{2}}^1 \sum_{\frac{a}{k} \geq c} \binom{k}{a} \left[t^a (1-t)^{k-a} + t^{k-a} (1-t)^a \right] dF(t)} \end{aligned} \quad /20/$$

Lemma 1:

If $\frac{\alpha_i}{\beta_i} (i = 1, 2, \dots; \alpha_i \geq 0, \beta_i > 0)$ is a monotone decreasing function of i then

$$\frac{\sum_{i=0}^n \alpha_i}{\sum_{i=0}^n \beta_i} \text{ is a monotone decreasing function of } n.$$

Proof:

From $\frac{a}{b} \leq \frac{c}{d}$ when $a, c \geq 0; b, d > 0$ follows

$$\frac{a}{b} \leq \frac{a+c}{b+d} \leq \frac{c}{d}$$

This can be shown as:

$$\begin{aligned} ad &\leq bc \\ ad + ab &\leq bc + ab \end{aligned}$$

$$\frac{a}{b} \leq \frac{a+c}{b+d}$$

Similarly can be shown the other side of the inequality. Now the proof can be completed by full induction on n .

In order to apply our lemma in (20) we make the substitution $i = k - a$:

/21/

$$\frac{\sum_{i=0}^{k-1-[kc]} \int_{\frac{1}{2}}^1 t \binom{k}{i} [t^i (1-t)^{k-i} + t^{k-i} (1-t)^i] dF(t)}{\sum_{i=0}^{k-1-[kc]} \int_{\frac{1}{2}}^1 \binom{k}{i} [t^i (1-t)^{k-i} + t^{k-i} (1-t)^i] dF(t)}$$

where $[kc]$ denotes the integer part of $k.c$.

Clearly decreasing c the number of terms increases without

surpassing the value $\left[\frac{k}{2}\right]$. So we have to show that the expression

/22/

$$\frac{\int_{\frac{1}{2}}^1 t \binom{k}{i} \left[t^i (1-t)^{k-i} + t^{k-i} (1-t)^i \right] dF(t)}{\int_{\frac{1}{2}}^1 \binom{k}{i} \left[t^i (1-t)^{k-i} + t^{k-i} (1-t)^i \right] dF(t)}$$

is a monotone decreasing function of i . Proving this we will need the following.

Lemma 2:

If for positive $f_1(x), f_2(x)$ on the interval $0 < x < 1$ there exists the integrals $\int_{\frac{1}{2}}^1 f_1(x) dF(x)$, $\int_{\frac{1}{2}}^1 f_2(x) dF(x)$, $\int_{\frac{1}{2}}^1 x f_1(x) dF(x)$, $\int_{\frac{1}{2}}^1 x f_2(x) dF(x)$ and $\frac{f_1(x)}{f_2(x)}$ is monotone decreasing then

/23/

$$\frac{\int_{\frac{1}{2}}^1 x f_1(x) dF(x)}{\int_{\frac{1}{2}}^1 f_1(x) dF(x)} < \frac{\int_{\frac{1}{2}}^1 x f_2(x) dF(x)}{\int_{\frac{1}{2}}^1 f_2(x) dF(x)}$$

holds.

Proof:

/24/

$$\int_{\frac{1}{2}}^1 \frac{f_1(x)}{\int_{\frac{1}{2}}^1 f_1(x) dF(x)} dF(x) = \int_{\frac{1}{2}}^1 \frac{f_2(x)}{\int_{\frac{1}{2}}^1 f_2(x) dF(x)} dF(x) = 1$$

so there exists such x for which

$$\frac{f_1(x)}{\int_{\frac{1}{2}}^1 f_1(x) dF(x)} > \frac{f_2(x)}{\int_{\frac{1}{2}}^1 f_2(x) dF(x)}$$

/25/

and there must be such x that the opposite inequality holds.

Because of monotonicity of $\frac{f_1(x)}{f_2(x)}$ there exists such $0 < p < 1$ that for $\frac{1}{2} < x < p$ inequality (25) holds and for $p < x < \frac{1}{2}$ the opposite is valid.

So we obtain:

$$\int_{\frac{1}{2}}^1 x \left(\frac{f_2(x)}{\int_{\frac{1}{2}}^1 f_2(x) dF(x)} - \frac{f_1(x)}{\int_{\frac{1}{2}}^1 f_1(x) dF(x)} \right) dF(x) > \int_{\frac{1}{2}}^1 p \left(\frac{f_2(x)}{\int_{\frac{1}{2}}^1 f_2(x) dF(x)} - \frac{f_1(x)}{\int_{\frac{1}{2}}^1 f_1(x) dF(x)} \right) dF(x) + \int_p^{\frac{1}{2}} \left(\frac{f_2(x)}{\int_{\frac{1}{2}}^1 f_2(x) dF(x)} - \frac{f_1(x)}{\int_{\frac{1}{2}}^1 f_1(x) dF(x)} \right) dF(x) = 0 \quad /26/$$

Q.E.D.

In order to apply our Lemma 2 for proving the monotonicity of (22) we have to show that

$$M = \frac{t^{i+1} (1-t)^{k-i-1} + t^{k-i-1} (1-t)^{i+1}}{t^i (1-t)^{k-i} + t^{k-i} (1-t)^i} \quad /27/$$

is monotone decreasing in $\frac{1}{2} < t < 1$

$$M = \frac{\frac{1}{1-t} + \left(\frac{t}{1-t}\right)^{k-2i-1}}{1 + \left(\frac{t}{1-t}\right)^{k-2i}} \quad /28/$$

substituting $x = \frac{t}{1-t}$ which is monotone increasing from 1 to ∞ :

$$M = \frac{x + x^{k-2i-1}}{1 + x^{k-2i}} \quad /29/$$

Considering the numerator of the derivative of M :

$$1 - x^{2k-4i-2} + (k-2i-1) \cdot x^{k-2i-2} \cdot (1-x^2) \quad /30/$$

where $1 < x < \infty$

and (30) is easily seen to be negative.

This completes our proof since M is monotone decreasing function of x and therefore of t as well. Thus Lemma 2 can be applied and after application of Lemma 1 our assertion can be proved.

3. Experimental investigations.

The searching method was applied in a modified way to improve the original Bledsoe-Browning method. We selected as matter of fact not functions but p -tuples at random which means 2^p functions at once and the sum of $\hat{h}(\varphi_i)$ for these functions were compared with a threshold constant.

This method was applied to recognition of small handwritten characters in a 8×8 receptor space for $p = 4$ /Fig. 1./. The sample volume was 200 characters and the examination was carried out by 400 independent characters. We found that until selecting 30 p -tuples having the threshold $c < 0,85$ the confidence of recognition is increasing function of c . The Figure 2. shows this fact where k denotes the number of selected good p -tuples and r the rate of correct recognition. When $c > 0,85$ the rate of correct recognition decreases. Selecting more than 30 p -tuples the rate of recognition increased only when there was an upper threshold as well ($0,75 < \hat{h}(\varphi_i) < 0,85$).

The plausible explanation of this fact is that the functions with large $\hat{h}(\varphi_i)$ represent the special biases of the sample not being representative for the problem itself. On the other hand as it was shown in the theoretical investigation the efficient functions in assamble carry less information than the less efficient ones.

In order to investigate the original searching method described above we selected the generated at random functions from the class of boolean functions defined by the Bledsoe-Browning method /minterms/. The task was to discriminate between two artificially given categories. There were two vectors given in the 10-dimensional measurement space:

$$\underline{X}^A = (0,7; 0,7; 0,7; 0,7; 0,7; 0,7; 0,7; 0,7; 0,7; 0,7)$$

$$\underline{X}^B = (0,3; 0,3; 0,3; 0,3; 0,3; 0,7; 0,7; 0,7; 0,7; 0,7)$$

which were considered as centers of mass of the two categories A and B. Adding noise to these vectors we generated the bina-

ry elements of the pattern sets. If a coordinate was 0,7 then with probability 0,7 we altered this coordinate to 1 and with probability 0,3 to 0. If a coordinate happened to be 0,3 the procedure was the opposite. These random changes were performed independently from each other. The result was that a binary vector could belong to both of the categories and the theoretically computed optimal probability of correct decision was only 0,92.

In our case because of the biasedness of the generated learning set this figure was slightly higher: 0,94.

This task as it might be shown can be solved by a Perceptron-type recognizer in the sense we investigated the p -tuple algorithm above. So it can be solved by computing only the number of common elements $m(X, Y)$ in the learning set and the pattern to be classified (δ).

So here we can not hope to increase the rate of correct recognition by applying the searching technique and the only advantage we gain is the minimization of the number of required functions and consequently the number of terms in the discriminant functions /cf. Figure 3., where k is the number of good functions/.

Conclusions

The searching method suggested in this paper has as minimal advantage to minimize the required memory capacity by cutting the number of terms in the discriminant function.

If we can manage to set an upper threshold during the function selection procedure or to introduce a measure of function-efficiency which takes into account the dependency between the separate functions as well, the rate of correct decisions clearly increases. So we can conclude that this is a successful way to adjust the recognition technique to the specific relevant characteristics of the task to be solved.

References:

1. Bledsoe, W.W. and Browning, I.: Pattern Recognition and Reading by Machine
Proceedings of the Eastern Joint Computer Conference, December 1959.
2. Novikoff, A.: On Convergence Proofs for Perceptrons Proceedings of the Symposium on Mathematical Theory of Automata, New York, April 1962.
Polytechnic Press, Brooklyn, N.Y., 1963.
3. Barát, J. and Muszély, Gy.: A Pattern Recognition Algorithm with Searching Method
Acta Technica Academiae Scientiarum Hungaricae, Tomus 59 /3-4/ 1967.
4. Petrov, A.P.: Possibilities of the Perceptron /in Russian/
Technical Cybernetics, No.6. 1964.
5. Cover, T.M.: Geometrical and Statistical Properties of Linear Threshold Devices
Stanford Electronics Laboratories Technical Report 6107-1. May, 1964.

