Tema 5. Clasificadores K-NN

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco–Euskal Herriko Unibertsitatea

9.1 Introducción

En este tema vamos a estudiar un paradigma clasificatorio conocido como K-NN (K-*Nearest Neighbour*). La idea básica sobre la que se fundamenta este paradigma es que un nuevo caso se va a clasificar en la clase más frecuente a la que pertenecen sus K vecinos más cercanos. El paradigma se fundamenta por tanto en una idea muy simple e intuitiva, lo que unido a su fácil implementación hace que sea un paradigma clasificatorio muy extendido.

Después de introducir el algoritmo K-NN básico y presentar algunas variantes del mismo, en este tema se estudian métodos para la selección de prototipos.

9.2 El algoritmo K-NN básico

La notación a utilizar (véase la Figura 1) en este tema es la siguiente:

		X_1	$\ldots$	X_j	$\ldots$	X_n	C
$(\mathbf{x}_1, c_1)$	1	x_{11}	$\ldots$	x_{1j}	$\ldots$	x_{1n}	c_1
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(\mathbf{x}_i, c_i)$	i	x_{i1}	$\ldots$	x_{ij}	$\ldots$	x_{in}	c_i
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
$(\mathbf{x}_N, c_N)$	N	x_{N1}	$\ldots$	x_{Nj}	$\ldots$	x_{Nn}	c_N
$\mathbf{x}$	$N+1$	$x_{N+1,1}$	$\ldots$	$x_{N+1,j}$	$\ldots$	$X_{N+1,n}$	?

Figura 1: Notación para el paradigma K-NN

- D indica un fichero de N casos, cada uno de los cuales está caracterizado por n variables predictoras, $X_1, \ldots, X_n$ y una variable a predecir, la clase C.
- Los N casos se denotan por

$$\begin{aligned} (\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N) \quad & \text{donde} \\ \mathbf{x}_i = (x_{i,1} \ldots x_{i,n}) \quad & \text{para todo } i = 1, \ldots, N \\ c_i \in \{c^1, \ldots, c^m\} \quad & \text{para todo } i = 1, \ldots, N \end{aligned}$$

  $c^1, \ldots, c^m$ denotan los m posibles valores de la variable clase C.
- El nuevo caso que se pretende clasificar se denota por $\mathbf{x} = (x_1, \ldots, x_n)$.

En la Figura 2 se presenta un pseudocódigo para el clasificador K-NN básico. Tal y como puede observarse en el mismo, se calculan las distancias de todos los casos ya clasificados al nuevo caso, $\mathbf{x}$, que se pretende clasificar. Una vez seleccionados los K casos ya clasificados, $D_{\mathbf{x}}^k$ más cercanos al nuevo caso, $\mathbf{x}$, a éste se le asignará la

COMIENZO
  Entrada: $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)\}$
        $\mathbf{x} = (x_1, \ldots, x_n)$ nuevo caso a clasificar
  PARA todo objeto ya clasificado (x_i, c_i)
        calcular $d_i = d(\mathbf{x}_i, \mathbf{x})$
  Ordenar $d_i (i = 1, \ldots, N)$ en orden ascendente
  Quedarnos con los K casos $D_{\mathbf{x}}^K$ ya clasificados más cercanos a $\mathbf{x}$
  Asignar a $\mathbf{x}$ la clase más frecuente en $D_{\mathbf{x}}^K$
FIN

Figura 2: Pseudocódigo para el clasificador K-NN

clase (valor de la variable C) más frecuente de entre los K objetos, $D_{\mathbf{x}}^k$. La Figura 3 muestra de manera gráfica un ejemplo de lo anterior. Tal y como puede verse en

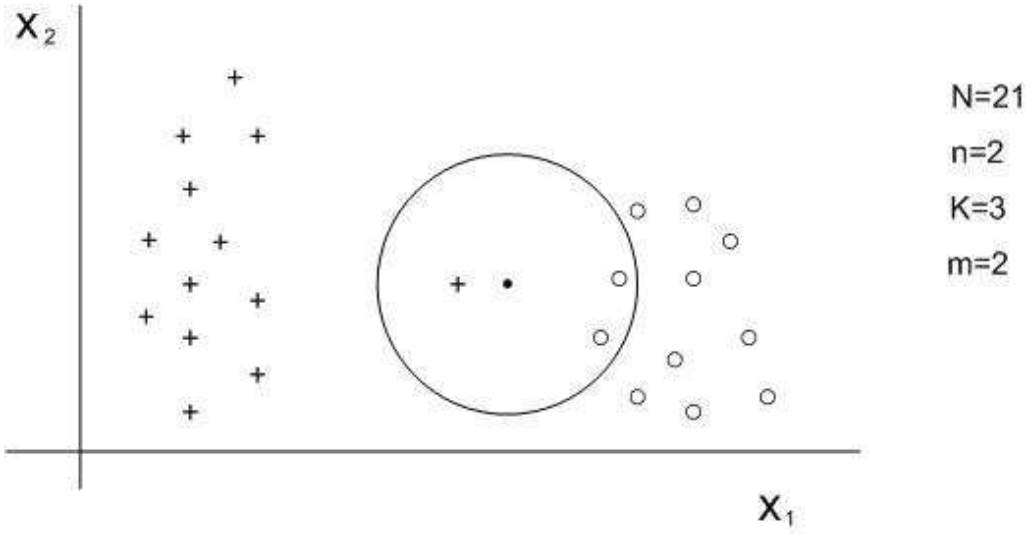

Figura 3: Ejemplo de aplicación del algoritmo K-NN básico

la Figura 3 tenemos 24 casos ya clasificados en dos posibles valores ($m = 2$). Las variables predictoras son X_1 y X_2, y se ha seleccionado $K = 3$. De los 3 casos ya clasificados que se encuentran más cercanos al nuevo caso a clasificar, $\mathbf{x}$ (representado por $\bullet$), dos de ellos pertenecen a la clase $\circ$, por tanto el clasificador 3-NN predice la clase $\circ$ para el nuevo caso. Nótese que el caso más cercano a $\mathbf{x}$ pertenece a la clase $+$. Es decir, que si hubiésemos utilizado un clasificador 1-NN, $\mathbf{x}$ se hubiese asignado a $+$.

Conviene aclarar que el paradigma K-NN es un tanto atípico si lo comparamos con el resto de paradigmas clasificatorios, ya que mientras que en el resto de paradigmas la clasificación de un nuevo caso se lleva a cabo a partir de dos tareas, como son la *inducción* del modelo clasificatorio y la posterior *deducción* (o aplicación) sobre el nuevo caso, en el paradigma K-NN al no existir modelo explícito, las dos tareas anteriores se encuentran colapsadas en lo que se acostumbra a denominar *transinducción*.

En caso de que se produzca un *empate* entre dos o más clases, conviene tener una *regla*

heurística para su ruptura. Ejemplos de reglas heurísticas para la ruptura de empates pueden ser: seleccionar la clase que contenta al vecino más próximo, seleccionar la clase con distancia media menor, etc.

Otra cuestión importante es la determinación del valor de K. Se constata empíricamente que el porcentaje de casos bien clasificados es no monótono con respecto de K (véase Figura 4), siendo una buena elección valores de K comprendidos entre 3 y 7.

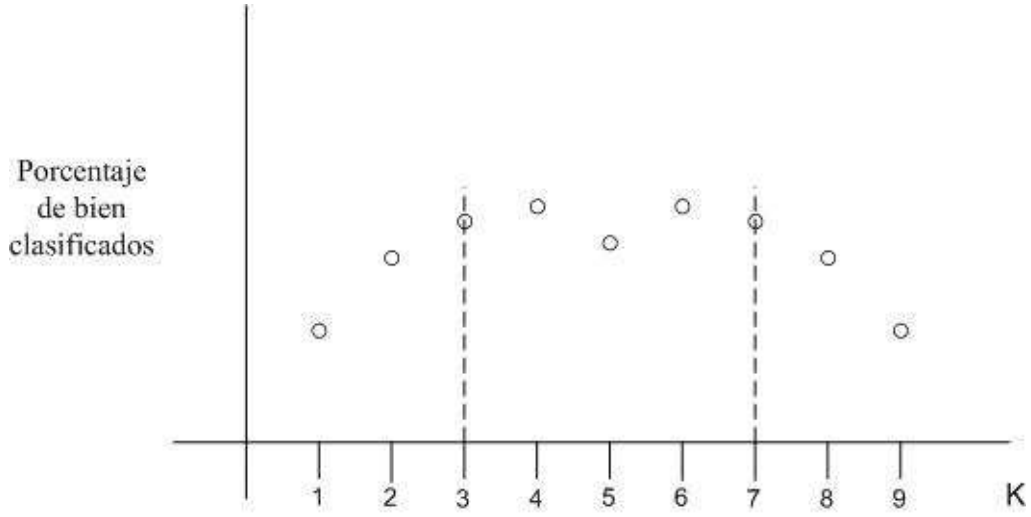

Figura 4: Ejemplo de la no monotocidad del porcentaje de bien clasificados en función de K

9.3 Variantes sobre el algoritmo básico

En este apartado vamos a introducir algunas variantes sobre el algoritmo básico.

9.3.1 K-NN con rechazo

La idea subyacente al K-NN con rechazo es que para poder clasificar un caso debo de tener ciertas garantías. Es por ello por lo que puede ocurrir que un caso quede sin clasificar, si no existen ciertas garantías de que la clase a asignar sea la correcta.

Dos ejemplos utilizados para llevar a cabo clasificaciones con garantías son los siguientes:

- el número de votos obtenidos por la clase deberá superar un *umbral prefijado*. Si suponemos que trabajamos con $K = 10$, y $m = 2$, dicho umbral puede establecerse en 6.
- establecimiento de algún tipo de *mayoría absoluta* para la clase a asignar. Así, si suponemos que $K = 20$, $m = 4$, podemos convenir en que la asignación del nuevo caso a una clase sólo se llevará a cabo en el caso de que la diferencia entre las frecuencias mayor y segunda mayor supere 3.

9.3.2 K-NN con distancia media

En el K-NN con distancia media la idea es asignar un nuevo caso a la clase cuya distancia media sea menor. Así que en el ejemplo de la Figura 5, a pesar de que 5 de los 7 casos más cercanos al mismo pertenecen a la clase $\circ$, el nuevo caso se clasifica como $+$, ya que la distancia media a los dos casos $+$ es menor que la distancia media a los cinco casos $\circ$.

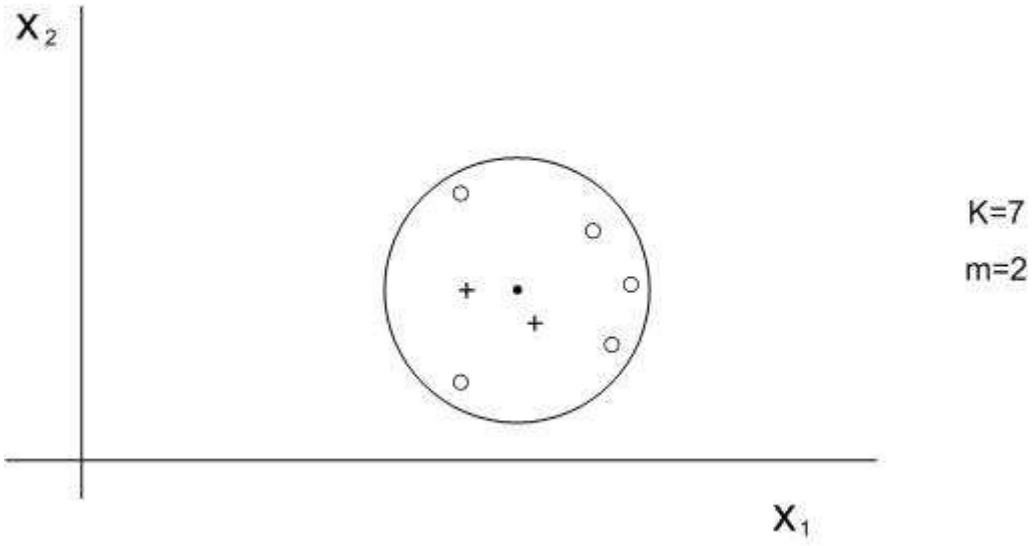

Figura 5: Ejemplo de ilustración del K-NN con distancia media

9.3.3 K-NN con distancia mínima

En el K-NN con distancia mínima se comienza seleccionando un caso por clase, normalmente el caso más cercano al baricentro de todos los elementos de dicha clase. En este paso se reduce la dimensión del fichero de casos a almacenar de N a m. A continuación se asigna el nuevo caso a la clase cuyo representante esté más cercano.

El procedimiento anterior puede verse como un 1-NN aplicado a un conjunto de m casos (uno por cada clase). El coste computacional de este procedimiento es inferior al K-NN genérico, si bien su efectividad está condicionada a la homogeneidad dentro de las clases (cuanto mayor sea dicha homogeneidad, se espera que el procedimiento sea más efectivo).

9.3.4 K-NN con pesado de casos seleccionados

La idea en el K-NN con el que se efectúa un pesado de los casos seleccionados es que los K casos seleccionados no se contabilicen de igual forma, sino que se tenga en cuenta la distancia de cada caso seleccionado al nuevo caso que pretendemos seleccionar. Como ejemplo podemos convenir en pesar cada caso seleccionado de manera inversamente proporcional a la distancia del mismo al nuevo caso. En la Figura 7 se puede consultar el peso w_i que se asignaría a cada uno de los 6 casos seleccionados provenientes de la Figura 6. Como resulta que de los 6 casos seleccionados, los pesos relativos a los casos tipo $\circ$, suman 2, cantidad inferior al peso de los dos casos tipo $+ \left(\frac{1}{0{,}7} + \frac{1}{0{,}8}\right)$, el K-NN con pesado de casos seleccionados clasificaría el nuevo caso como perteneciente a la clase $+$.

9.3.5 K-NN con pesado de variables

En todas las aproximaciones presentadas hasta el momento, la distancia entre el nuevo caso que se pretende clasificar, $\mathbf{x}$, y cada uno de los casos $\mathbf{x}_r$, $r = 1, \ldots, N$ ya clasificados pertenecientes al fichero de casos D, da el mismo peso a todas y cada una de las n variables, $X_1, \ldots, X_n$. Es decir, la distancia $d(\mathbf{x}, \mathbf{x}_r)$ entre $\mathbf{x}$ y $\mathbf{x}_r$ se calcula

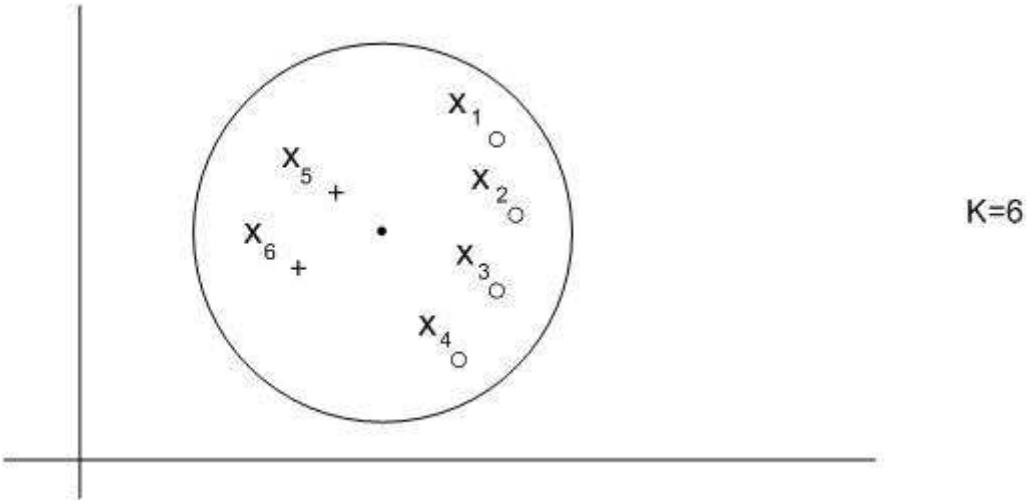

Figura 6: Ejemplo de ilustración del K-NN con pesado de casos seleccionados

	$d(\mathbf{x}_i, \mathbf{x})$	w_i
$\mathbf{x}_1$	2	0,5
$\mathbf{x}_2$	2	0,5
$\mathbf{x}_3$	2	0,5
$\mathbf{x}_4$	2	0,5
$\mathbf{x}_5$	0,7	1/0,7
$\mathbf{x}_6$	0,8	1/0,8

Figura 7: Peso a asignar a cada uno de los 6 objetos seleccionados

por ejemplo por medio de la distancia euclídea, de la siguiente manera:

$$d(\mathbf{x}, \mathbf{x}_r) = \sum_{j=1}^{n} (x_j, x_{rj})^2$$

Esta manera de calcular la distancia, es decir, otorgando la misma importancia a todas las variables, puede resultar peligrosa para el paradigma K-NN en el caso de que algunas de las variables sean irrelevantes para la variable clase C. Este es el motivo por el que resulta interesante el utilizar distancias entre casos que ponderen cada variable de una manera adecuada. Es decir, la distancia entre $\mathbf{x}$ y $\mathbf{x}_r$ se calcularía:

$$d(\mathbf{x}, \mathbf{x}_r) = \sum_{j=1}^{n} w_j (x_j, x_{rj})^2$$

con lo cual la variable X_j tiene asociado un peso w_j que habrá que determinar. Para clarificar la idea supongamos los datos de la Figura 8. En dicha figura se puede observar que la variable X_1 es irrelevante para la variable C, ya que las 6 veces en las que X_1 toma el valor 0, en tres de ellas C toma el valor 1, y en las otras tres el valor 1, mientras que de las 6 veces en las que X_1 toma el valor 1, en tres de ellas C toma el valor 0, y las tres restantes toma el valor 1. Esta situación es totalmente opuesta a la situación con la variable X_2. De ls 6 veces en que X_2 toma el valor 0, en 5 casos C vale 1, valiendo 0 en el caso restante, mientras que de las 6 veces en las que X_2 toma el valor 1, en 5 casos C vale 0 valiendo 1 en el caso restante.

Ante la situación anterior parece intuitivo que en el cálculo de las distancias se deba

X_1	X_2	C
0	0	1
0	0	1
0	0	1
1	0	1
1	0	1
1	1	1
0	1	0
0	1	0
0	1	0
1	1	0
1	1	0
1	0	0

Figura 8: La variable X_1 no es relevante para C, mientras que la variables X_2 si lo es

de pesar más la aportación de la variable X_2 que la aportación de la variable X_1.

Una manera de calcular la ponderación w_i de cada variable X_i es a partir de la medida de información mínima $I(X_i, C)$ entre dicha variable X_i y la variable clase C, definida de la siguiente manera:

$$I(X_i, C) = \sum_{x_i, c} p_{(X_i,C)}(x_i, c) \log \frac{p_{(X_i,C)}(x_i, c)}{p_{X_i}(x_i) \cdot p_C(c)}$$

para todo $i = 1 \ldots, n$.

Esta medida de información mutua entre dos variables se interpreta como la reducción en la incertidumbre sobre una de las variables cuando se conoce el valor de la otra variable. Cuanto mayor sea la medida de información mutua entre dos variables mayor será la "dependencia.ªistente entre las mismas.

Para el asunto que nos concierne en este apartado, el peso w_i asociado a la variable X_i será proporcional a la medida de información mutua $I(X_i, C)$ entre X_i y C. Con los datos de la Figura 8 calculamos $I(X_1, C)$ y $I(X_2, C)$

$$I(X_1, C) = p_{(X_1,C)}(0,0) \log \frac{p_{(X_1,C)}(0,0)}{p_{X_1}(0) \cdot p_C(0)} + p_{(X_1,C)}(0,1) \log \frac{p_{(X_1,C)}(0,1)}{p_{X_1}(0) \cdot p_C(1)} +$$

$$p_{(X_1,C)}(1,0) \log \frac{p_{(X_1,C)}(1,0)}{p_{X_1}(1) \cdot p_C(0)} + p_{(X_1,C)}(1,1) \log \frac{p_{(X_1,C)}(1,1)}{p_{X_1}(1) \cdot p_C(1)} =$$

$$\frac{3}{12} log \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{3}{12} log \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{3}{12} log \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{3}{12} log \frac{\frac{3}{12}}{\frac{6}{12} \cdot \frac{6}{12}} = 0$$

$$I(X_2, C) = p_{(X_2,C)}(0,0) \log \frac{p_{(X_2,C)}(0,0)}{p_{X_2}(0) \cdot p_C(0)} + p_{(X_2,C)}(0,1) \log \frac{p_{(X_2,C)}(0,1)}{p_{X_2}(0) \cdot p_C(1)} +$$

$$p_{(X_2,C)}(1,0) \log \frac{p_{(X_2,C)}(1,0)}{p_{X_2}(1) \cdot p_C(0)} + p_{(X_2,C)}(1,1) \log \frac{p_{(X_2,C)}(1,1)}{p_{X_2}(1) \cdot p_C(1)} =$$

$$\frac{1}{12} log \frac{\frac{1}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{5}{12} log \frac{\frac{5}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{5}{12} log \frac{\frac{5}{12}}{\frac{6}{12} \cdot \frac{6}{12}} + \frac{1}{12} log \frac{\frac{1}{12}}{\frac{6}{12} \cdot \frac{6}{12}}$$

9.4 Reducción del fichero de casos inicial

En este apartado veremos dos aproximaciones distintas, con las que se trata de paliar la fuerte dependencia del costo computacional del clasificador K-NN con respecto del número de casos, N, almacenados. Los distintos métodos de reducción del fichero de casos inicial los podemos clasificar en *técnicas de edición* y en *técnicas de condensación*.

En las técnicas de edición, el subconjunto del fichero inicial se obtiene por medio de la eliminación de algunos de los casos, mientras que las técnicas de condensación tratan de obtener el subconjunto del fichero inicial por medio de la selección de unos cuantos casos.

9.4.1 Edición de Wilson

La idea en la edición propuesta por Wilson es el someter a prueba a cada uno de los elementos del fichero de casos inicial. Para ello, para cada caso se compara su clase verdadera con la que propone un clasificador K-NN obtenido con todos los casos excepto el mismo. En el caso de que ambas clases no coincidan, el caso es eliminado. En cierto modo, el método tiene una analogía con la validación *leave-one-out*.

Se puede también considerar una edición de Wilson repetitiva, y parar el procedimiento cuando en 2 selecciones sucesivas no se produzcan cambios.

9.4.2 Condensación de Hart

El condensado de Hart efectúa una selección de casos que pueden considerarse raros.

Para ello, para cada caso, y siguiendo el orden en el que se encuentran almacenados los casos en el fichero, se construye un clasificador K-NN con tan sólo los casos anteriores al caso en cuestión. Si el caso tiene un valor de la clase distinto al que le asignaría el clasificador K-NN, el caso es seleccionado. Si por el contrario la clase verdadera del caso coincide con la propuesta por el clasificador K-NN, el caso no se selecciona.

Es claro que el método de condensación de Hart es dependiente del orden en que se encuentren almacenados los casos en el fichero.

Referencias

1. D. Aha, D. Kibler, M.K. Albert (1991). Instance-based learning algorithms, *Machine Learning* **6**, 37-66

2. S. Cost, S. Salzberg (1993). A weighted nearest neighbour algorithm for learning with symbolic features, *Machine Learning* **10(1)**, 57-78

3. B.V. Dasarathy (1991). *Nearest Neighbour (NN) Norms: NN Pattern Recognition Techniques*, IEEE Computer Society Press

4. E. Fix, J.L.Hodges Jr. (1951). Discriminatory analysis, nonparametric discrimination. *Project 21-49-004, Rept. 4*, USAF School of Aviation Medicine, Randolf Field

5. P.E. Hart (1968). The condensed nearest neighbour rule, *IEEE Transactions on Information Theory*, **IT-14**, 515-516

6. D.L. Wilson (1972). Asymptotic properties of nearest neighbour rules using edited data. *IEEE Transactions on Systems, Man and Cybernetics*, Vol 2, 408-421