CLASIFICACIÓN

NO SUPERVISADA

CLUSTERING

Y

MAPAS AUTOORGANIZATIVOS

(KOHONEN)

(RECUPERACIÓN Y ORGANIZACIÓN DE LA INFORMACIÓN)

Cristina Díaz Moreno info@clustering.50webs.com

Indice

Introducción

Uno de los principales problemas a los que se enfrenta la sociedad de la información, en la actualidad, es la gestión óptima y productiva de la documentación disponible. En otras palabras, diariamente se generan grandes cantidades de datos y es imprescindible establecer técnicas que nos ayuden a localizar, lo antes posible, la **información** que nos resulta relevante según nuestras necesidades. En resumen, es necesaria una correcta **organización de la información** para que su **recuperación** sea lo más completa posible.

Es en este punto donde entran en juego los sistemas de resumen automático de documentos, empleados para optimizar el tratamiento (obtención, filtrado, clasificado y extracción) de la información (en cualquier idioma), a fin de poder proporcionar al usuario, de forma eficaz y eficiente, exclusivamente los datos que precisa.

Éste es un problema tradicional de Inteligencia Artificial en el ámbito del Aprendizaje Automático: la **Clasificación automática**.

Resumen del contenido

Las técnicas de **clasificación** automática se pueden agrupar inicialmente como supervisadas o **no supervisadas** y, aunque en esta página se va a hacer especial hincapié en las segundas, se proporciona una breve descripción de ambas.

Conocida la diferencia entre ambas técnicas es más fácil introducirnos en la descripción de los **clasificadores**, determinando a qué dominios se aplican en la actualidad, que parámetros lo configuran y que tipo son los más usados.

Además en esta Web se dispone de información sobre dos de los **clasificadores no supervisados** más conocidos y empleados en la **organización (y recuperación) de la información**: **k-medias** y los **mapas auto-organizativos de Kohonen**.

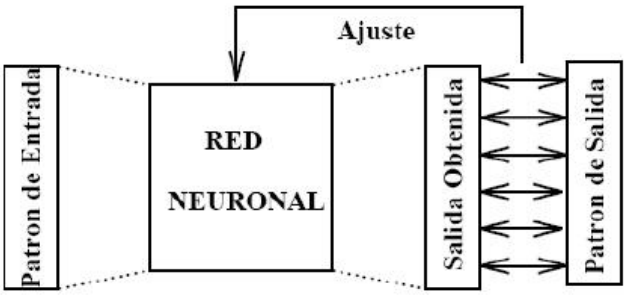
Por último se establece la relación entre los conocimientos teóricos listados anteriormente y la labor que realizan los buscadores para poder dar los resultados que más se ajusten a nuestras necesidades al realiza una consulta.

Aprendizaje (Supervisado VS no supervisado)

A continuación se detalla una breve descripción de ambos tipos de **aprendizaje**, más adelante se comenta el principal problema en el entrenamiento de redes de neuronas artificiales: el sobreajuste.

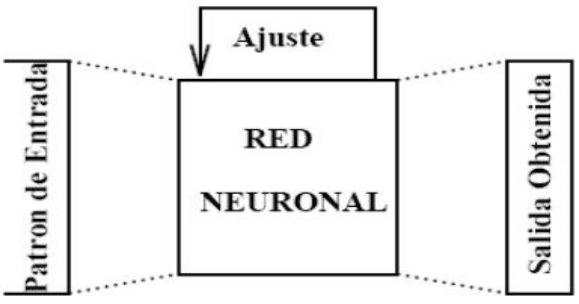
Aprendizaje supervisado

- necesita un profesor que mida el funcionamiento del sistema
- maneja información de error o de control
- esta información se emplea para guiar al sistema. Hay varios algoritmos que establecen cómo se realiza esta retroalimentación, el más conocido o empleado es el *backtracking*



Aprendizaje no supervisado

- no utiliza información externa
- reajuste automático de los parámetros
- **autoorganización de la información**



Aprendizaje y entrenamiento

Cuando se está entrenando una red es imprescindible establecer una condición de parada óptima que minimice el error pero evitando, en todo momento, el sobreajuste (overfitting). que se produce cuando una red es incapaz de generalizar para casos nuevos.

Para evitar que esto ocurra se recomienda dividir el conjunto de ejemplos disponibles:

- *Conjunto de entrenamiento*: usado para ajustar el valor de los pesos de la red
- *Conjunto de validación*: usado para medir la eficacia de la red. Debe ser:
    - *Significativo* (debe contener ejemplos pertenecientes a todas las clases establecidas)
    - *Representativo* (debe guardar la relación existente entre los ejemplos del conjunto de entrenamiento)

Clasificación

Algunos de los dominios en los que la clasificación automática se emplea:

- Visión artificial (reconocimiento de caras)
- Reconocimiento de caracteres
- **Clasificación de documentos**
- Reconocimiento del habla

Una de las principales características de la clasificación automática, que la hacen tan atractiva para la **recuperación y organización de la información** de los documentos, es su rapidez y su capacidad de síntesis de datos relevantes en la toma de decisiones.

División en las técnicas de clasificación

- Clasificación
  - A partir del conocimiento de la existencia de un conjunto de clases, determinar la regla para asignar cada nueva observación (o ejemplo) a la clase que pertenece
  - Determina reglas de asignación a clases conocidas
  - Aprendizaje supervisado
- **Agrupamiento (clustering)**
  - A partir de una serie de observaciones determina si existen clases en las que dichas observaciones puedan ser agrupadas
  - Determinar la existencia de clases en las que agrupar (número y características de las clases desconocidas a priori)
  - Aprendizaje no supervisado

Parámetros

La elección del tipo de clasificador viene supeditada tanto al dominio del problema a tratar como a los parámetros que tengan más relevancia según dicho dominio. En el caso de la **recuperación de la información** se consideran primordiales los dos primeros.

- **Calidad** (capacidad de acierto de la regla o del clasificador. Errores de clasificación: falso positivo y falso negativo)
- **Velocidad** (velocidad de respuesta crítica aunque se pierda calidad)

- Explicabilidad (información sobre qué está ocurriendo con el clasificador y el por qué de la aplicación de ciertas operaciones)
- Tiempo de aprendizaje (en entornos cambiantes es necesario modificar las reglas de funcionamiento)

Tipos de clasificadores

- Discriminantes lineales / no lineales
  - Dividen el espacio de estados en regiones (definidas por el corte de hiperplanos) teniendo cuidado de establecer una clase por región
  - Por ejemplo: *Discriminantes lineales, discriminantes logísticos, discriminantes cuadráticos, redes de neuronas*
- Métodos basados en reglas
  - Dividen el espacio de estados de forma recursiva estableciendo dos bloques a partir de cada atributo
  - Cada bloque puede ser subdividido con la ayuda de otro atributo
  - Proceso repetido hasta que no mejora la clasificación estableciéndose una regla por atributo
  - La unión de reglas define el clasificador
  - Por ejemplo: *ID3, AC2, Cal5, CN2, C4.5, CART, Árboles de Bayes, Regla IT*
- Métodos de estimación de densidades
  - Fijan para cada región del espacio la probabilidad de que un elemento situado en ella pertenezca a una clase
  - Clasificación por vecindad (ante un nuevo patrón se le asigna la clase más probable en función de la distancia que le separe de los prototipos designados).
    - Por ejemplo: K-medias (no supervisada), *K-vecinos (supervisada), LVQ (supervisada)*
  - Funciones de base radial
  - Naive Bayes
  - Poliárboles
  - Mapas autoorganizativos de Kohonen

K-medias

Clasificación por vecindad no supervisada

Parte de patrones (observaciones) sin etiquetar y un número de prototipos definido (por el diseñador del clasificador).

- El espacio de entrada se divide en k clases y k prototipos
- Mueve los prototipos una vez ha realizado el aprendizaje con todas las observaciones

Entrenamiento

El objetivo de este algoritmo es intentar situar los prototipos de forma tal que aquellos patrones cercanos (distancia euclídea) sean similares entre sí.

- Minimizar distancia entre patrón y prototipo

$$J=\sum_{i=1}^{k}\sum_{n=1}^{m} M_{in}\,||x_n - A_i||^2$$

- Calcular si el patrón pertenece o no a un prototipo (1 si pertenece y 0 en caso contrario)

$$M_{in}=\begin{cases} 1 & \text{si } ||x_i-A_i||^2 < ||x_n-A_s||^2 \ \forall s\neq i,\ s=1,2,\ldots,k \\ 0 & \text{en caso contrario} \end{cases}$$

- Desplazamiento de los prototipos al centro formado por los patrones que representan

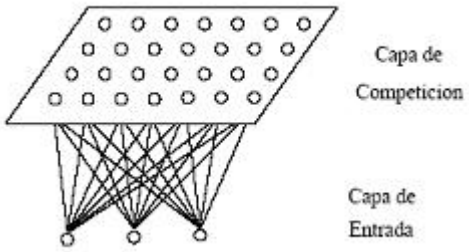
Validación

Ante un nuevo dato el sistema simplemente lo comparará con los prototipos y lo clasificará según la clase definida para el más cercano.

Entrenamiento potencialmente lento y clasificación rápida

Mapas autoorganizativos (Kohonen)

Clasificación no supervisada

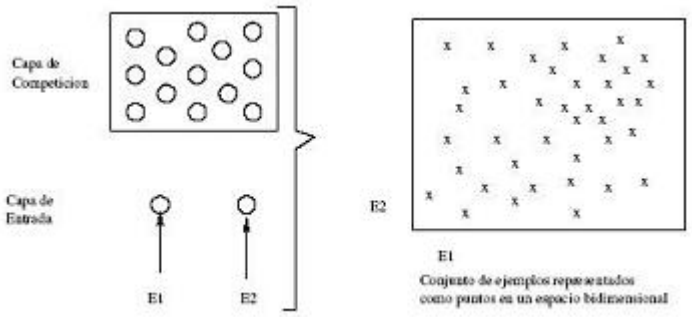


Redes de neuronas de dos capas:

- Cada neurona de competición es una categoría
- Cada neurona de entrada está conectada con cada una de las células de la capa de competición (células que se distribuyen inicialmente de forma aleatoria)
- Para cada ejemplo se calcula la salida de cada célula de competición y nos quedamos con la mejor (ganadora)

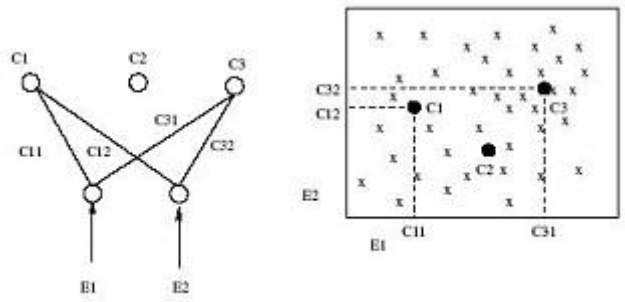
Procedimiento

1. Se recibe el ejemplo de entrada (n-dimensional)
   Los ejemplos son representables como puntos en un espacio n-dimensional.

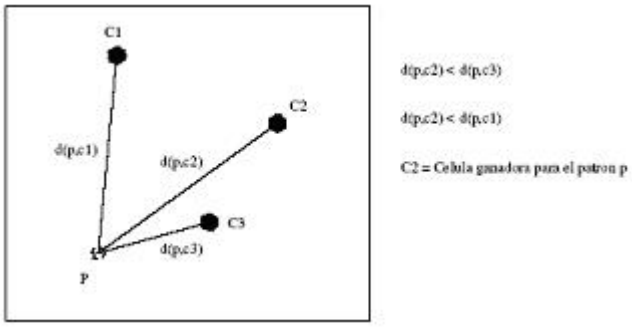


2. Se propaga por las conexiones hasta llegar a la capa de competición (competición que se realiza en base a un modelo de interacción lateral)
   Los prototipos también se pueden representar en el espacio y sus

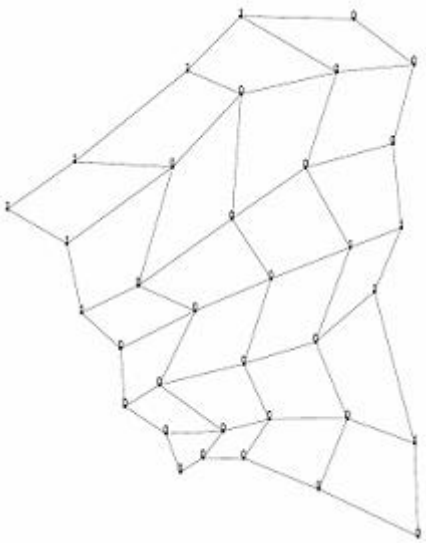
coordenadas quedan determinadas por los pesos de las neuronas de la capa de competición



3. Cada célula de esta capa de competición produce una salida al comparar el ejemplo con sus pesos
4. Se selecciona el prototipo cuya distancia al ejemplo sea menor (célula ganadora)



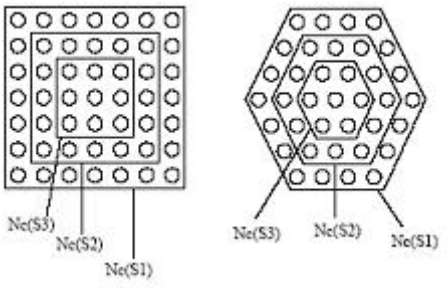
5. Los pesos de la célula ganadora se modifican para acercarse ligeramente al ejemplo modificando así el mapa de prototipos inicial



Vecindarios

Cuando se trata de los mapas autoorganizativos de Kohonen no sólo importa la distancia del ejemplo a los distintos prototipos existentes sino también la clase a la que pertenecen los ejemplos cercanos. Por esta razón es útil el concepto de vecindario en esta técnica.

Los vecindarios más empleados son los rectangulares o los hexagonales.



Características

- Los prototipos se desplazan según se van procesando los ejemplos (al contrario que en k-medias)
- El aprendizaje es más rápido y eficaz
- No requiere aprendizaje en el procesamiento de datos nuevos

Herramientas

En la siguiente página (no oficial) se puede descargar un conjunto de programas empleados para el aprendizaje de los mapas de Kohonen. Este conjunto de programas recibe el nombre de SOM_PAK.

- Neural Networks Research Centre (Public domain program packages)

Buscadores

Ahora vamos a centrarnos en el uso del **clustering** como técnica para, en base a la **recuperación y organización de la información**, obtener el máximo aprovechamiento de la información digital disponible.

Conceptos

- *Recuperación de información* (dado un conjunto amplio de datos se obtienen aquellos que cumplan determinados criterios -palabras clave-)
- *Extracción de información* (obtiene la información relevante de uno o varios documentos)
- *Clustering* (crea, de forma automática, clasificaciones de documentos a partir de similitudes en su contenido)
- *Cluster* (agrupación de elementos con características similares)
- *Resumen* (términos más utilizados dentro de un documento y/o similitudes entre varios documentos)

Técnicas para resumen (empleadas en clustering)

Técnicas utilizadas para obtener un resumen de uno o varios documentos ya sea por una búsqueda, aplicación de un filtro o simplemente la necesidad de **clasificarlo**.

Basadas en la superficie del texto

- No se realiza análisis lingüístico
- Trata el texto como una cadena de caracteres
- Clásicamente selecciona los términos estadísticamente más frecuentes en el documento
- Selecciona como resumen las oraciones con el mayor número de términos más frecuentes del documento
- La posición de los elementos en el texto (títulos, párrafos...) también es relevante

Basadas en los términos del texto

- Reconocimiento y clasificación del léxico utilizado
- Permite reconocer unidades lingüísticas (nombre, verbo...)
- Emplea analizadores morfológicos y desambiguadores léxicos

- Establece relaciones entre términos (semánticas y temáticas)

Basadas en la estructura del discurso

- Requieren algún tipo de tratamiento estructural del documento
- Detecta los fragmentos del discurso más relevantes

Para poder crear los **clusters**, los documentos se representan como vectores de términos (cuyo tamaño es igual al del vocabulario del conjunto recuperado tras el análisis del documento).

Clustering y buscadores

Debido al gran volumen de información que hay que procesar, junto con la eficacia y eficiencia solicitadas a los buscadores, el uso de *técnicas de clustering* ha supuesto una gran mejora en los resultados proporcionados por los buscadores.

Los tres primeros (*Vivisimo, Clusty y iBoogie*) usan exclusivamente agrupación para mostrar los documentos en categorías según la cantidad de términos que coinciden en sus textos.

Kartoo en cambio no sólo se centra en la creación de clusters, a partir de la documentación disponible y en base a la búsqueda realizada, sino que también representa gráficamente los resultados obtenidos en forma de mapa.

Para probar el funcionamiento de los 4 buscadores (cómo realizan y presentan los resultados) vamos a realizar una búsqueda guiada *"clasificacion no supervisada"*.

Como se puede apreciar todos los buscadores realizan agrupamientos de términos similares si bien es cierto que la representación de resultados de Kartoo resulta más espectacular.

- Vivisimo (ahora es Clusty)
- Clusty
- iBoogie
- Kartoo


Vivísimo
search.vivisimo.com
Clustered Results
clasificacion no supervisada (147)
Clase (8)
Reconocimiento, Patrones (9)
Investigacion (7)
Algoritmos (9)
Teledetección (7)
Mapas autoorganizativos (6)
Usos del suelo (6)
Redes Neuronales (6)
Nacional, Universidad (5)
Espectral, Análisis (4)
More
Find in clusters:
Enter Keywords
Clusty
web news imac
clasificacion no su
clusters sources sites
All Results (150)
Imágenes (32)
Clase (8)
Reconocimiento, Patrones (9)
Investigacion (7)
Algoritmos (9)
Mapas autoorganizativos (6)
Usos del suelo (6)
Redes Neuronales (6)
Nacional, Universidad (5)
Teledetección (5)
more | all clusters
find in clusters:
Find
iBoogie BETA
Web
clasific
All results
Supervisada de imágenes
Análisis
Datos
Revista
Reconocimiento de patrones
Información
Imagenes
Clasificación no supervisada por métodos
Sistemas de clasificación
Proceso de clasificación
Report
Producto
Suelo
Template
Técnicas
Problema de clasificación
Classification
Nacional
Realizó una clasificación no supervisada
Programas
More...


KartOO
www.kartoo.com
help
Clasificacion no supervisada
Search
english pages
options
Products
Saved map
New map
Topics :
procesamiento digital
clasificación digital
para clasificacion supervis
para la clasificación super
clasificación supervisada p
áreas de entrenamiento
métodos de clasificación
algoritmos genéticos
seleccion de variables
clasificación no supervisad
entre
grupo
agrupada
multispec
para
computing
metodos
digital
proceso
imaging
informes
algorithm
clustering
www.uncosa.org
math.uprm.edu
clustering
proceso
para
informes
caece.edu.ar
entre
www.freewebs.com
imaging
clustering 50webs.com
coc.inaoep.mx
extracción
agrupada
multispec
nationalzoo.si.edu
www.umbc.edu
www.globe.gov
digital
algorithm
grupo
computing
www.sc.ehu.es
metodos
www.utem.cl
chiapas.mediosindependientes.or
print the map
Send a map
Add a site
Add a Topic
save the map...
Preview by
THUMBSHOTS
KartOO
next map
10 200 Found results
1 - 15

Referencias

Páginas que me han sido muy útiles no sólo para documentar el contenido del sitio sino también para la estructura, estilo y posicionamiento del mismo.

- Scalab (grupos de investigación del departamento de informática de la Universidad Carlos III)
- Sistemas de resumen automático de documentos
- Browsing y clustering