TRATAMIENTO DIGITAL DE SEÑALES
Ingeniería de Telecomunicación (4º, 2º c)

Unidad 10ª: Agrupamiento

Aníbal R. Figueiras Vidal
Jesús Cid Sueiro
Ángel Navia Vázquez

Área de Teoría de la Señal y Comunicaciones
Universidad Carlos III de Madrid

Agrupamiento (“clustering”)

En ocasiones, no se puede realizar una clasificación supervisada (p.ej., los datos no están etiquetados), o no se desea hacerlo (directamente).

Puede interesar **agruparlos** según algún criterio; normalmente

- para optimizar un **objetivo**
- de acuerdo con una **similitud** (implica **minimizar** la similitud global) (o disimilitud: ídem maximizar)

con posibilidad de satisfacer requisitos secundarios (restricciones, estructura,...)

Así se procede de forma **no supervisada**: aunque también puede decirse que hay una “supervisión integrada” según el principio seguido: cuya elección ha de ser adecuada para que el agrupamiento tenga utilidad.

El agrupamiento puede servir:

* como un **paso exploratorio** para una posterior clasificación:
  - obteniendo información útil para ésta, p.ej., qué variables son relevantes;
  - permitiendo proponer hipótesis o clases, e incluso una primera verificación de su validez;

* para una **reducción de datos**: pasando a trabajar con un representante de cada grupo (o varios); así:
  - se reduce la dificultad de un primer diseño de un clasificador; e incluso puede ganarse en generalización;
  - se facilita un (razonable) proceso de etiquetado (por grupos): hacerlo para todas las muestras podría ser muy costoso o imposible;

* como un paso para **simplificar un proceso auxiliar** en la resolución del problema de que se trate; p.ej.,
  - para realizar imputaciones de valores perdidos de forma local;
  - para reducir la carga computacional de un proceso de selección de muestras (las significativas para encontrar la solución);

* a fin de convertir un **problema global en un conjunto de problemas locales**: permitiendo proponer un diseño distinto para cada grupo, o según la relación de la muestra tratada con los representantes de los grupos.

(Naturalmente, este proceso es subóptimo: ya que el agrupamiento no se hace -en principio- para optimizar el subsiguiente tratamiento).

Agrupamiento para optimizar un objetivo

Los procedimientos más difundidos son los basados en considerar las muestras como generadas por una mezcla de densidades:

$$p(\mathbf{x}) = \sum_{i=1}^{G} Pr(G_i)\, p(\mathbf{x} \mid G_i)$$

siendo G el número de grupos de la mezcla y $\{G_i\}$ dichos grupos. Se resuelven mediante la aplicación del criterio de Máxima Verosimilitud.

Posibles agrupaciones posteriores pueden hacerse determinísticamente, de acuerdo con el máximo, o probabilísticamente, proporcionalmente a las $Pr(G_j \mid \mathbf{x})$.

Otras optimizaciones, como la del error cuadrático total

$$\sum_{j=1}^{G} \sum_{\mathbf{x}^{(k)} \in G_j} \left\| \mathbf{x}^{(k)} - \mathbf{m}_j \right\|_2^2$$

dan lugar a planteamientos comunes con los métodos que se basan en una similitud/disimilitud.

Agrupamiento por similitud/disimilitud

Volvamos, como ejemplo inicial, al caso de proceder con la distancia euclídea como medida de disimilitud.

Es obvio que:

– si los grupos estuviesen dados, sus representantes serían las medias grupales

$$\mathbf{m}_j = \frac{1}{\#G_j} \sum_{\forall \mathbf{x}^{(k)} \in G_j} \mathbf{x}^{(k)}$$

– si las medias grupales estuviesen dadas, se podría agrupar por distancias directamente

$$j^* = \arg\left[\min_j \left\{ \left\| \mathbf{x}^{(k)} - \mathbf{m}_j \right\|_2^2 \right\}\right] \quad \Rightarrow \quad \mathbf{x}^{(k)} \in G_{j^*}$$

pero hay que resolver ambas cosas a la vez.

Se hace (básicamente) mediante algoritmos secuenciales o algoritmos paralelo (bloque iterados).

Algoritmos secuenciales

Su esquema general es:

- inicializar
- pasar muestra a muestra
  - asignando según la (di)similitud establecida
  - actualizando los representantes según las asignaciones
- detener por resultado aceptable o saturación

Naturalmente, presentan problemas por mínimos locales.

El **Algoritmo (Secuencial) Básico** sigue esta forma: teniendo en cuenta cuál es la verdadera variación del error cuadrático por el traslado de una muestra de un grupo a otro;

– para el grupo, G_j, al que llega la muestra, $\mathbf{x}^{(k)}$, la nueva media será:

$$\mathbf{m}'_j = \frac{1}{M_j+1}\left(\sum_{x^{(i)}\in G_j}\mathbf{x}^{(i)} + \mathbf{x}^{(k)}\right) = \frac{M_j}{M_j+1}\mathbf{m}_j + \frac{1}{M_j+1}\mathbf{x}^{(k)}$$

siendo M_j el número (previo) de muestras en el grupo, $\mathbf{x}^{(k)}$ la muestra incorporada, y $\mathbf{m}_j$ la media anterior; de modo que

$$\mathbf{x}^{(i)} - \mathbf{m}'_j = \mathbf{x}^{(i)} - \mathbf{m}_j - \frac{1}{M_j+1}\left(\mathbf{x}^{(k)} - \mathbf{m}_j\right)$$

con lo que el error cuadrático para la muestra $\mathbf{x}^{(i)} \in G_j$ es

$$\left\| \mathbf{x}^{(i)} - \mathbf{m}'_j \right\|_2^2 = \left\| \mathbf{x}^{(i)} - \mathbf{m}_j \right\|_2^2 - 2\frac{1}{M_j+1}\left(\mathbf{x}^{(i)} - \mathbf{m}_j\right)^T\left(\mathbf{x}^{(k)} - \mathbf{m}_j\right) + \frac{1}{\left(M_j+1\right)^2}\left\| \mathbf{x}^{(k)} - \mathbf{m}_j \right\|_2^2$$

sumando para todas las $\mathbf{x}^{(i)} \in G_j$, el segundo sumando se anula $\left(\sum\left(\mathbf{x}^{(i)} - \mathbf{m}_j\right) = \mathbf{0}\right)$, y el tercero multiplicado por M_j da el incremento del error;

además se incrementa el error en

$$\left\|\mathbf{x}^{(k)}-\mathbf{m}'_j\right\|_2^2=\left\|\mathbf{x}^{(k)}-\frac{M_j}{M_j+1}\mathbf{m}_j-\frac{1}{M_j+1}\mathbf{x}^{(k)}\right\|_2^2=\frac{M_j^2}{(M_j+1)^2}\left\|\mathbf{x}^{(k)}-\mathbf{m}_j\right\|_2^2$$

con lo que el incremento total es:

$$\left[\frac{M_j}{(M_j+1)^2}+\frac{M_j^2}{(M_j+1)^2}\right]\left\|\mathbf{x}^{(k)}-\mathbf{m}_j\right\|_2^2=\frac{M_j}{M_j+1}\left\|\mathbf{x}^{(k)}-\mathbf{m}_j\right\|_2^2$$

– análogamente, el grupo que pierde la muestra reduce su error cuadrático en

$$\frac{M_{j'}}{M_{j'}-1}\left\|\mathbf{x}^{(k)}-\mathbf{m}_{j'}\right\|_2^2$$

con lo que el Algoritmo (Secuencial) Básico es como se presenta a continuación.

Algoritmo (Secuencial) Básico:

1. Inicializar: establecer $\{\mathbf{x}^{(k)} \in G_j\}$ y fijar $\mathbf{m}_j$= promedio$\{\mathbf{x}^{(k)} \in G_j\}$
2. Tomar una muestra $\mathbf{x}^{(k)}$; sea
   - si M_j=1, volver a 2;
   - si M_j>1, calcular

$$\text{decr}_{j'} = \frac{M_{j'}}{M_{j'}-1}\left\|\mathbf{x}^{(k)}-\mathbf{m}_{j'}\right\|_2^2 \qquad \text{incr}_j = \frac{M_j}{M_j+1}\left\|\mathbf{x}^{(k)}-\mathbf{m}_j\right\|_2^2, \quad j \neq j'$$

y, si incr_{j^*} es mínimo, asignar a G_{j^*} la muestra y recalcular las medias:

$$\mathbf{m}_{j^*} \rightarrow \frac{M_{j^*}}{M_{j^*}+1}\mathbf{m}_{j^*} + \frac{1}{M_{j^*}+1}\mathbf{x}^{(k)}$$

$$\mathbf{m}_{j'} \rightarrow \frac{M_{j'}}{M_{j'}-1}\mathbf{m}_{j'} - \frac{1}{M_{j'}-1}\mathbf{x}^{(k)}$$

3. Volver a 2 hasta que se cumpla el criterio de parada

La asignación inicial puede hacerse mediante una etapa del algoritmo K-medias, que veremos a continuación. Una asignación completamente aleatoria puede plantear problemas de mínimos locales.

Algoritmos paralelo

Tienen como esquema general:

- inicializar
- pasar muestra a muestra la totalidad de ellas
  asignando según la (di)similitud establecida
- actualizar los representantes según las asignaciones anteriores
- detener por resultado aceptable o saturación.

Naturalmente, también presentan problemas por mínimos locales.

El bien conocido **Algoritmo K-medias** (C-medias, ISODATA básico) asigna por comparación directa de distancias. Procede como se indica a continuación.

Algoritmo K-medias:

1. Inicializar: fijar $\{\mathbf{m}_j\}$
2. Para todas las muestras $\mathbf{x}^{(k)}$

$$j^* = \arg\left[\min_j \left\{ \left\| \mathbf{x}^{(k)} - \mathbf{m}_j \right\|_2^2 \right\} \right] \quad \Rightarrow \quad \mathbf{x}^{(k)} \in G_{j^*}$$

3. Recalcular las medias:

$$\mathbf{m}_j = \frac{1}{M_j} \sum_{\mathbf{x}^{(k)} \in G_j} \mathbf{x}^{(k)}$$

4. Volver a 2 hasta que se cumpla el criterio de parada.

Sobre el número de grupos

Los algoritmos anteriores lo dan por supuesto: naturalmente, no tiene por qué ser así.

Puede determinarse extendiendo los algoritmos a:

- formas **crecientes**: aumentando el número de grupos por **división** de los existentes o por **creación** de grupos nuevos;
- formas **autoconstructivas**: completando el proceso de crecimiento por **fusión** de grupos existentes.

(No son aconsejables las formas decrecientes, por motivos computacionales).

Siguen algunos ejemplos clásicos.

Algoritmo LBG (Linde-Buzo-Gray)

1. Fijar G=2
2. Ejecutar el Algoritmo K-medias
3. Mientras no se alcance el criterio de parada (resultado aceptable o saturación o $G_{máx}$), duplicar G y volver a 2.

(se duplica para poder indexar los grupos en binario).

Un modo habitual de duplicar es dividir las medias del modo:

$$\mathbf{m}_j^+ = \mathbf{m}_j + \alpha\sigma_j\mathbf{u}$$

$$\mathbf{m}_j^- = \mathbf{m}_j - \alpha\sigma_j\mathbf{u}$$

donde **u** es un vector unitario aleatorio, σ_j es la desviación típica del grupo, y α un escalón seleccionable entre 0 y 1.

ISODATA ("**I**terative **S**elf-**O**rganizing **D**ata **A**nalysis **T**echnique **A**lgorithm)

Es un algoritmo que incluye (versión "standard"):

- división de grupos: a la que se procede cuando
  - su varianza muestral rebasa un cierto umbral, y
  - la distancia promedio de sus muestras al representante es la mayor de todas o el número actual de grupos es menor que un cierto valor (normalmente, la mitad del número deseado o esperado)

- fusión de grupos: fundiendo dos si su distancia entre representantes es menor que un cierto umbral; procediendo (en caso de varias candidaturas) en orden de distancias crecientes y hasta un número máximo de fusiones en cada paso.

Algoritmo de Hall

Es un algoritmo secuencial que incluye la creación de grupos mediante el manejo de un umbral θ:

1. Se crea la primera clase con una muestra, $(\mathbf{x}^{(1)})$;
2. Se presenta la muestra $\mathbf{x}^{(k)}$ y se elige el grupo G_j más cercano:
   - si $\left\|\mathbf{x}^{(k)} - \mathbf{m}_j\right\|_2^2 < \theta$ o si $G = G_{máx}$, $\mathbf{x}^{(k)} \in G_j$
   - si $\left\|\mathbf{x}^{(k)} - \mathbf{m}_j\right\|_2^2 > \theta$ y $G<G_{máx}$, se crea una nueva clase con $\mathbf{x}^{(k)}$
3. Si es necesario, se actualizan las representaciones de los grupos;
4. Se termina al terminarse las muestras.

Obviamente, es defecto de este algoritmo la asignación de muestras a grupos antes de que éstos estén totalmente definidos.

Versión **Modificada**: para reducir el problema,

- parte de las muestras se utilizan como se ha visto;
- el resto se asigna directamente.

Otro inconveniente (común con todos los secuenciales) es la dependencia del orden de presentación de las muestras.

Versión con **Dos Umbrales**: para reducir la dependencia,

- para asignar a un grupo se utiliza θ_1
- para crear un nuevo grupo, $\theta_2 > \theta_1$

y, en los casos intermedios, se **demora** la asignación de la muestra a una etapa posterior.

El parámetro θ (o el par θ_1,θ_2) tiene una importante utilidad: como resulta $G=G(\theta)$, se pueden explorar resultados para varios valores de θ, con varias ordenaciones de los datos para cada valor; eligiendo el G más frecuente para cada θ, se puede adoptar como θ más conveniente el valor que esté en el centro de la zona plana más extensa (elección para mayor robustez).

Todos los algoritmos de error cuadrático tienen versiones generalizadas inmediatas: en que la distancia euclídea se sustituye por una medida (general) de disimilitud $d(\mathbf{x}^{(k)},\mathbf{c}_j)$, en que $\mathbf{c}_j$ es un **centroide**

$$\mathbf{c}_j : \min_{\mathbf{c}_j} \sum_{\mathbf{x}^{(k)} \in G_j} d\left(\mathbf{x}^{(k)}, \mathbf{c}_j\right)$$

y obviamente se agrupa por mínima disimilitud; minimizando así la disimilitud total

$$D = \sum_{j=1}^{G} \sum_{\mathbf{x}^{(k)} \in G_j} d\left(\mathbf{x}^{(k)}, \mathbf{c}_j\right)$$

Si d es una medida de **distorsión**, se obtiene lo que se llama un **Cuantificador Vectorial** (VQ, "Vector Quantizer"): representar las $\mathbf{x}_i$ por sus $\mathbf{c}_j$ (y, con un buen diseño, lo mismo para muestras **x** de la población representada por $\{\mathbf{x}_i\}$) da lugar a una **distorsión** (media) mínima; con la ventaja de que, a efectos de transmisión, los $\mathbf{c}_j$ se pueden sustituir por índices binarios (de ahí la duplicación del Algoritmo LBG). Los algoritmos paralelo aplicados a este fin (K-medias modificado, LBG modificado) se suelen denominar **Algoritmos de Lloyd Generalizados** (Lloyd fue el diseñador del cuantificador escalar óptimo).

Trabajo: Influencia de los errores de canal en los VQ; codificación fuente más canal.

Agrupamiento mixto

Puede ser interesante agrupar según una similitud, pero elegir como representante otro distinto del centroide correspondiente a ésta: p.ej., si se utiliza el agrupamiento como vía de reducción de datos, puede convenir separar ambas cosas.

Un ejemplo relevante es el **Algoritmo de Kohonen**, que consiste en clasificar de acuerdo con el producto escalar y actualizar por acercamiento directo.

Es decir: los representantes $\{\mathbf{w}_j\}$ compiten para quedarse con las muestras $\{\mathbf{x}^{(k)}\}$ de acuerdo con

$$j^* = \arg\left[\max_j \left\{ \mathbf{w}_j^T \mathbf{x}^{(k)} \right\} \right] \quad \Rightarrow \quad \mathbf{x}^{(k)} \in G_{j^*}$$

y la actualización se hace según

$$\mathbf{w}_j(k+1) = \begin{cases} \mathbf{w}_j(k) + \eta(k) \left[\mathbf{x}^{(k)} - \mathbf{w}_j(k) \right], & j = j^* \\ \mathbf{w}_j(k), & j \neq j^* \end{cases}$$

(que se llama modo **acretivo** o WTA: “Winner Takes All”; los modos **interpolativos** actualizan también representantes perdedores, en menor grado).

$\eta(k)$ empieza con un valor cercano a 1 y también se hace tender a 0.

El tipo de actualización (aprendizaje) que aquí aparece por primera vez, en forma no supervisada, se denomina **Hebbiano**: **refuerza la capacidad de ganar** (cualitativamente, en este caso) del representante vencedor ante casos semejantes (nótese que, al incluir $\eta\ \mathbf{x}^{(k)}$ en $\mathbf{w}_j(k+1)$, el producto escalar de éste con muestras que se parezcan a $\mathbf{x}^{(k)}$ tenderá a crecer).

Obviamente, esta forma de agrupamiento no minimiza el error cuadrático: ya que la victoria se produce según un producto escalar, y, aunque se "equilibra" la competición, los grupos establecidos no lo serán para minimizar el error cuadrático (salvo que todas las muestras tengan igual módulo).

Como, obviamente, dentro de cada grupo el representante $\mathbf{w}_j$ toma como valor al acabar el entrenamiento la media de las muestras del grupo (ya que va incorporando valores de las muestras capturadas y olvidando su propio pasado), lo que se minimiza es el error cuadrático para una asignación dada; es decir, en total

$$\sum_{\forall \mathbf{x}^{(k)}} \sum_{j=1}^{G} \text{ind}_j\left(\mathbf{x}^{(k)}\right) \left\| \mathbf{x}^{(k)} - \mathbf{w}_j \right\|_2^2$$

siendo $\text{ind}_j(\mathbf{x}^{(k)})$ el indicador de pertenencia de $\mathbf{x}^{(k)}$ a G_j (grupo de $\mathbf{w}_j$) : con el entrenamiento acretivo visto, vale 1 ó 0.

Si se desea un entrenamiento global, pueden utilizarse indicadores de pertenencia suavizados (derivables) $\{z_j(\mathbf{x}^{(k)})\}$ (con suma unitaria), lo que equivale a un funcionamiento interpolativo (p.ej., según el valor de productos escalares); y se puede aplicar gradiente para el entrenamiento

$$\mathbf{w}_j^{(k+1)} = \mathbf{w}_j^{(k)} + \eta(k) \frac{\partial\left(\left\{ z_j\left(\mathbf{x}^{(k)}\right) \right\} \left\| \mathbf{x}^{(k)} - \mathbf{w}_j^{(k)} \right\|_2^2 \right)}{\partial \mathbf{w}_j^{(k)}} \quad , \forall j$$

Inicialización

Se ha mencionado que los algoritmos que preceden son sensibles a problemas de mínimos locales. La sensibilidad es alta por tratarse de algoritmos de carácter **competitivo**: los representantes se disputan las muestras, y se aproximan a ellas cuando ganan; pues bien; cualquier desequilibrio en la competición llevará a una solución no deseada.

Evidentemente, tales desequilibrios se darán con facilidad si la inicialización es pobre.

Utilizar como representantes iniciales G muestras tomadas al azar suele ser suficiente para los algoritmos paralelo, ya que la actualización de los representantes tiene lugar después de una competición entre todas las muestras.

En el caso de los algoritmos secuenciales, la dificultad es mayor: un representante que gane y se acerque a una zona con muchas muestras puede dominarla, y reducir la actividad de los otros representantes. Para evitarlo, hay que "equilibrar" la competición en su fase inicial; pudiendo emplearse:

* los métodos tradicionales de **modificación** de muestras, como
  - añadir a las muestras un ruido (que las dispersa) de nivel decreciente;
  - utilizar el método de **refuerzo radial**, trabajando con

    $\beta(k)\,\mathbf{x}^{(k)}$

    con $\beta(0)$ bajo y creciente hacia 1 con k;

que tienen efectos correctores limitados y ralentizan el agrupamiento;

* los **métodos de arrastre**, que acercan a la muestra en cuestión tanto el centroide ganador como los de un entorno alrededor de él, éstos con pasos decrecientes a lo largo del proceso (también puede disminuir el entorno);

* los **métodos de penalización**, que dificultan que los centroides que ganan repetidamente lo sigan haciendo; como los procedimientos "**sensibles a la frecuencia**" (FS, "Frequency Sensitive"), que clasifican de acuerdo con la disimilitud multiplicada por el número previo de victorias de cada centroide (o una función creciente de este número).

Los dos últimos tipos ofrecen buenas prestaciones.