CIMPA-UCR

Clasificación Binaria

Índices de agregación

Aplicación de metaheurísticas

Caso binario

- $\Omega = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \{0,1\}^p$
- Definir un criterio (aditivo) $W(P)$:

$$\min_{P \in \mathbf{P}_k} W(P) = \sum_{l=1}^{k} \delta(C_l)$$

donde $\mathbf{P}_k$ es el conjunto de todas las particiones de Ω en k clases y δ mide la homogeneidad de las clases C_l

Criterios de homogeneidad

$$\delta_{\min}(C) = \min_{x,x' \in C}\{d(x,x')\}$$

$$\delta_{\max}(C) = \max_{x,x' \in C}\{d(x,x')\}$$

$$\delta_{\text{sum}}(C) = \sum_{x,x' \in C} d(x,x')$$

$$\delta_{\text{pon}}(C) = \frac{1}{2|C|}\sum_{x,x' \in C} d(x,x')$$

Criterios de homogeneidad (2)

$$\delta_{\text{med}}(C) = \frac{1}{|C|(|C|-1)} \sum_{x,x' \in C} d(x,x')$$

$$\delta_{\text{var}}(C) = \frac{1}{|C|(|C|-1)} \sum_{x,x' \in C} [d(x,x') - \mu(C)]^2$$

$$\delta_{L_1}(C) = \sum_{x \in C} \left\| x - m(C) \right\|_1 = \sum_{j=1}^{p} \min(a_j, |C| - a_j)$$

Propiedades

- Todos los criterios $\delta_{\min},\ldots,\delta_{L1}$ tienen la propiedad de monotonicidad
- Óptimo para $\delta_{\min}$: k-1 clases unitarias
- Existe un óptimo para $\delta_{\min},\ldots,\delta_{L1}$ con clases no vacías
- Cualquier óptimo para δ_{sum}, δ_{pon} y δ_{L1} tiene clases no vacías
- δ_{var} satisface una propiedad de Huygens

Propiedad tipo Huygens

Para cualquier clase C y cualquier número real β se satisface la descomposición:

$$\frac{1}{|C|(|C|-1)} \sum_{x,x' \in C} [d(x,x') - \beta]^2 = \delta_{\text{var}}(C) + [\mu(C) - \beta]^2$$

Fórm. de recurrencia para $\delta(C_j - \{x_i\})$

$$\delta_{\text{sum}}(C_j - \{x_i\}) = \delta_{\text{sum}}(C_j) - \sum_{x \in C_j} d(x, x_i)$$

$$\delta_{\text{pon}}(C_j - \{x_i\}) = \frac{n_j}{n_j - 1}\delta_{\text{pon}}(C_j) - \frac{1}{n_j - 1}\sum_{x \in C_j} d(x, x_i)$$

$$\delta_{\text{med}}(C_j - \{x_i\}) = \frac{n_j}{n_j - 2}\delta_{\text{med}}(C_j) - \frac{2}{(n_j - 1)(n_j - 2)}\sum_{x \in C_j} d(x, x_i)$$

$$\delta_{\text{var}}(C_j - \{x_i\}) = \frac{n_j}{n_j - 2}\delta_{\text{var}}(C_j) - [\mu(C_j - \{x_i\}) - \mu]^2 - \frac{2}{(n_j - 1)(n_j - 2)}\sum_{x \in C_j - \{x_i\}} [d(x, x_i) - \mu]^2$$

Fórm. de recurrencia para $\delta(C_l \cup \{x_i\})$

$$\delta_{\text{sum}}(C_l \cup \{x_i\}) = \delta_{\text{sum}}(C_l) + \sum_{x \in C_l} d(x, x_i)$$

$$\delta_{\text{pon}}(C_l \cup \{x_i\}) = \frac{n_l}{n_l + 1} \delta_{\text{pon}}(C_l) + \frac{1}{n_l + 1} \sum_{x \in C_l} d(x, x_i)$$

$$\delta_{\text{med}}(C_l \cup \{x_i\}) = \frac{n_l - 1}{n_l + 1} \delta_{\text{med}}(C_l) + \frac{2}{n_l (n_l + 1)} \sum_{x \in C_l} d(x, x_i)$$

$$\delta_{\text{var}}(C_l \cup \{x_i\}) = \frac{n_l - 1}{n_l + 1} \delta_{\text{var}}(C_l) + \frac{n_l - 1}{n_l + 1} [\mu(C_l \cup \{x_i\}) - \mu]^2 + \frac{2}{n_l (n_l + 1)} \sum_{x \in C_l} [d(x, x_i) - \mu(C_l \cup \{x_i\})]^2$$

Fórm. actualización para L_1

$$\delta_{L_1}(C_j - \{x_i\}) = \sum_{r=1}^{p} \min(a_{jr} - x_{ir}, |C_j| - 1 - a_{jr} + x_{ir})$$

$$\delta_{L_1}(C_l \cup \{x_i\}) = \sum_{r=1}^{p} \min(a_{lr} + x_{ir}, |C_l| + 1 - a_{lr} - x_{ir})$$

$$a_{jr}^{new} = a_{jr}^{old} - x_{ir}$$

$$a_{lr}^{new} = a_{lr}^{old} + x_{ir}$$

CIMPA-UCR

Uso de Heurísticas de Optimización

- Classical methods find local optima of W
- We have used heuristics with good characteristics:
  1. Simulated annealing
  2. Tabu search
  3. Genetic algorithms

Uso de heurísticas

- Transferencias $C_j \xrightarrow{i} C_l$
- Calcular $\delta(C_j - \{x_i\})$ y $\delta(C_l \cup \{x_i\})$
- **Sobrecalentamiento simulado**: escoger al azar *i* y *l*, aplicar la regla de Metropolis
- **Búsqueda tabú**: generar una muestra de vecinos seleccionando *i* y *l*; escoger el mejor vecino no tabú (usando un criterio de aspiración)
- **Algoritmo genético**: "cromosomas" = particiones, selección, mutaciones, cruzamiento

CIMPA-UCR

Sob.Sim en Particionamiento

A partir de la partición P se genera P' así:

- Escoger al azar (unif. en [1,n]) un objeto $\mathbf{x} \in \Omega$
- Escoger al azar(unif. en [1,k]) un índice de clase l
- Colocar a **x** en la clase C_l

Nota: corresponde a lo que S. Régnier llamaba una **transfererencia**

Características

- **Reversibilidad**: la probabilidad de $P \rightarrow P'$ es la misma que la probabilidad de $P' \rightarrow P$
- **Connectivity**: siempre es posible generar cualquier partición P' a partir de cualquier P (hay un número finito de transferencias)
- Los vecindarios tienen el mismo tamaño: $n(k\text{-}1)$
- $G_{ss'} = 1/n(k\text{-}1)$

Particionamiento con BT

- Estado: partición P en k clases de Ω
- Criterio: minimizar W (se debe escoger δ)
- Movimiento: crear P' por la transferencia de un único elemento a una nueva clase
- Lista tabú: indicatriz de la clase que conten{ia al objeto que fue transferido
- Aplicamos un criterio de aspiración

CIMPA-UCR

AG para particionamiento

- Estados o cromosomas: índices de clases

$$(2\ \ 2\ \ 3\ \ 1\ \ 1\ \ 1\ \ 3\ \ 2)$$

$$x_1\ x_2\ x_3\ \ \ldots\ \ x_n$$

- Función de adaptación: $B(P) = W(\Omega) - W(P)$
- Selección: ruleta aleatoria proporcional a B
- Cruzamiento: con probabilidad p_c (cromos.)
- Mutación: con probabilidad p_m (alelos)

Datos de universidades alemanas

- 49 universidades alemanas (Späth)
- Tabla 49×56 (presencia-ausencia)
- Uso de δ_{sum} y δ_{pon}
- $k = 3$, $W = 287$ (SS=100%, BT = 80%, k-medias: 20 veces)
- $k = 6$, $W = 228$ (SS=76%, BT = 70%, k-medias: 20 veces con $W = 234$)

Resultados para δ_{sum}, δ_{pon}

- 20 aplicaciones de SS, BT y AG
- 100 ejecuciones de k-medias basado en transferencias
- *W* es el mejor valor encontrado y % el porcentaje de veces que este valor fue encontrado en las 20 aplicaciones de cada método.

CIMPA-UCR

Resultados para δ_{sum}, δ_{wsm}

Datos	k	SS		BT		AG		KM	
		W	%	W	%	W	%	W	%
Simulados	4	0.0	100	0.0	100	2.0	20	0.0	34
Späth	6	89.9	100	89.9	75			89.9	5
Pejibaye	6	263.9	100	263.9	100	744.4	10	334.6	1
Simulados	4	0.0	100	0.0	100	0.0	40	0.0	62
Späth	6	10.7	100	10.7	100	10.7	75	10.7	12
Pejibaye	6	18.2	100	18.2	95	25.4	10	18.2	2

Conclusiones en Clasif.Bin.

- El estudio muestra que es posible llevar a cabo particionamiento eficiente de datos binarios usando índices de agregación que no usan el concepto de centroide, combinados con metaheurísticas de optimización combinatoria.
- El desarrollo de propiedades teóricas de estos índices de agregación, que son importantes para la formulación de los algoritmos.
- Los resultados para δ_{sum}, δ_{pon} y δ_{L1} son los mejores índices.

Conclusiones en Clasif.Bin.

Las pruebas llevadas a cabo en algunos conjuntos de datos muestran que los métodos basados en metahaurísticas pueden dar mejores resultados que los métodos tradicionales.

Conclusiones

- Para tabla de tamaño medio y muchas clases, SS es mejor y más rápido que BT y AG.
- Hasta ahora, AG no ha dado buenos resultados pero los parámetros aún pueden ser afinados para mejorar los resultados.
- Estamos planeando una comparación sistemática (simulaciones, métodos,...).