Machine Learning

Unidad # 3 - Aprendizaje Supervisado Avanzado y Aprendizaje No Supervisado

CC57 – 2019-1

Profesor
Andrés Melgar

Competencias a adquirir en la sesión

- Al finalizar la sesión el alumno comprenderá el funcionamiento del **aprendizaje inductivo**.
- Al finalizar la sesión el alumno implementará **modelos algorítmos** usando algoritmos no supervisados.
- Al finalizar la sesión el alumno **entenderá** el algoritmo de **cobweb**.
- Al finalizar la sesión el alumno **aplicará** el algoritmo de **cobweb** para obtener modelos algorítmicos.

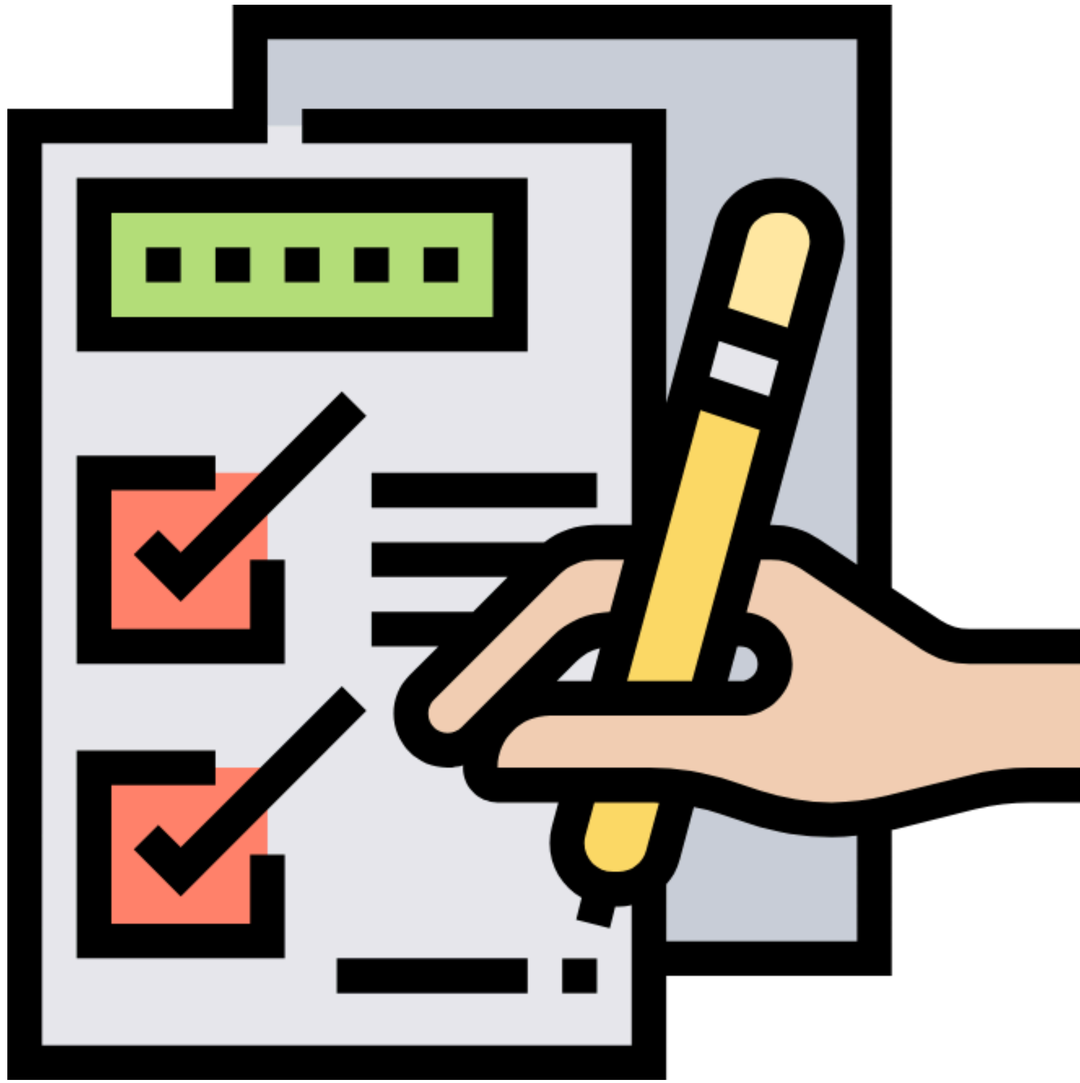


Imagen obtenida a través de la URL https://www.flaticon.com/free-icon/checklist_1467445

Métricas de Evaluación

Texto guía

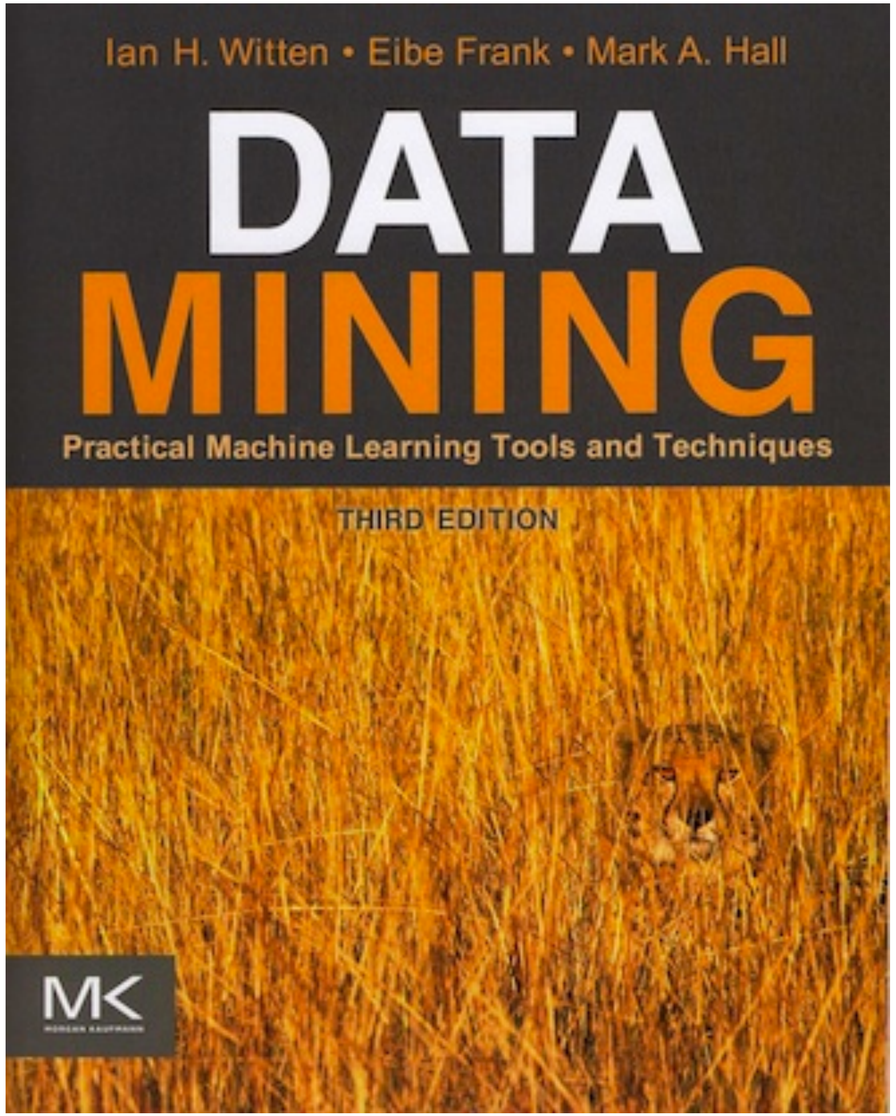
Witten, Ian H., Frank, Eibe, and Hall, Mark A.. 2011. ***Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations.*** San Francisco: Elsevier Science & Technology.

CHAPTER

Implementations: Real Machine Learning Schemes

6

6.8 CLUSTERING



Agrupamiento jerárquico

- La formación de un par inicial de grupos y luego considerar de forma recursiva si vale la pena dividirlos produce una jerarquía que se puede representar como un árbol binario llamado **dendrograma.**

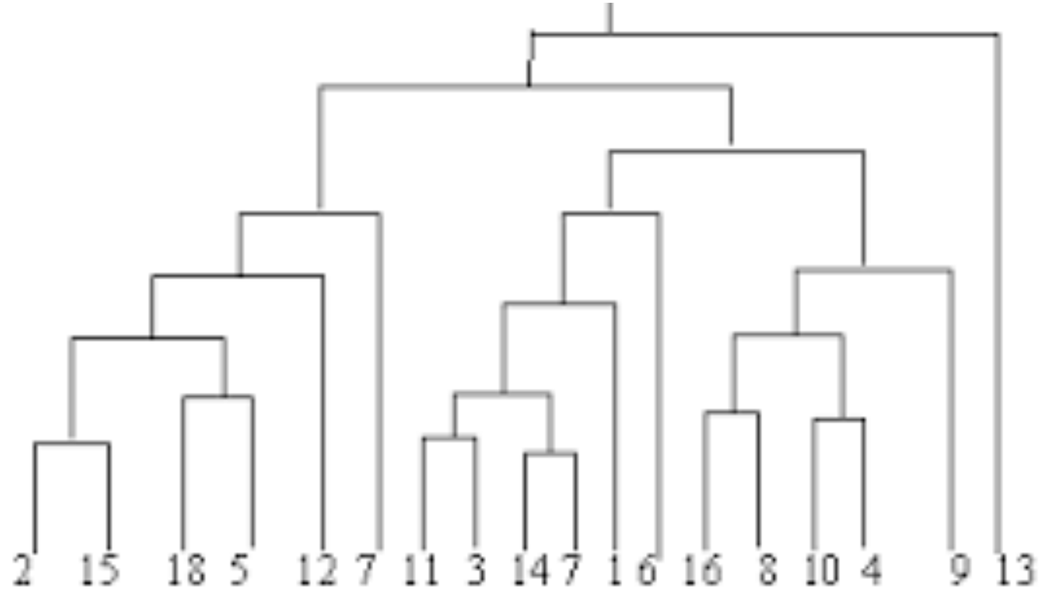
- La misma información podría ser representada como un diagrama de **Venn** de conjuntos y subconjuntos.

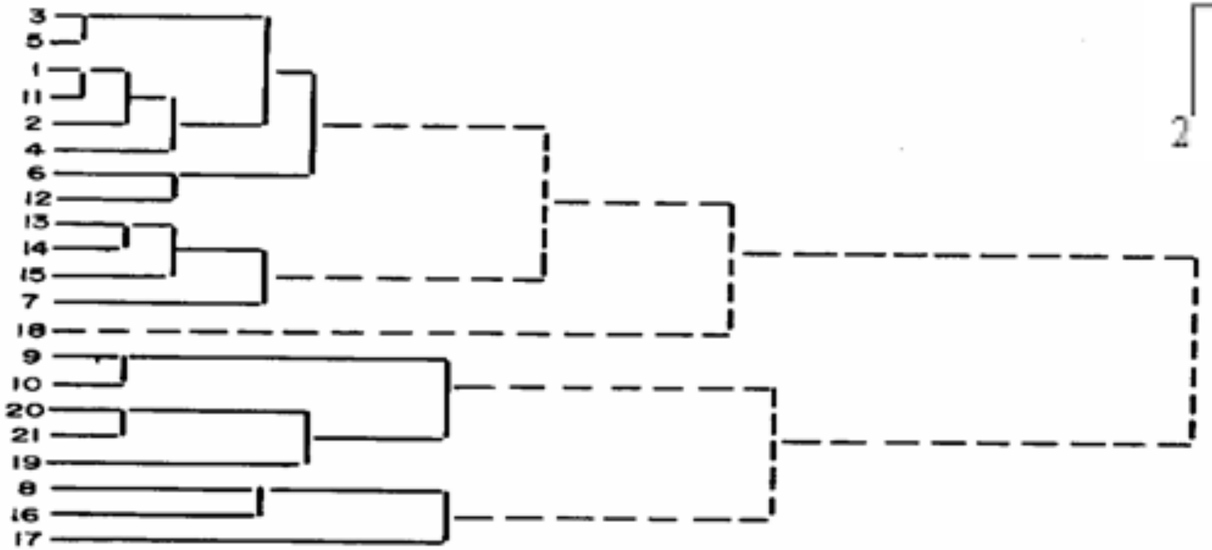
  - La restricción que la estructura es jerárquica corresponde al hecho de que, a pesar de subconjuntos pueden incluir uno del otro, **no pueden entrelazarse.**

  - En algunos casos existe una medida del **grado de disimilitud** entre los grupos en cada conjunto; a continuación, la altura de cada nodo en el dendrograma se puede hacer proporcional a la disimilitud entre sus hijos. Esto proporciona un diagrama de fácil interpretación de un agrupamiento jerárquico.

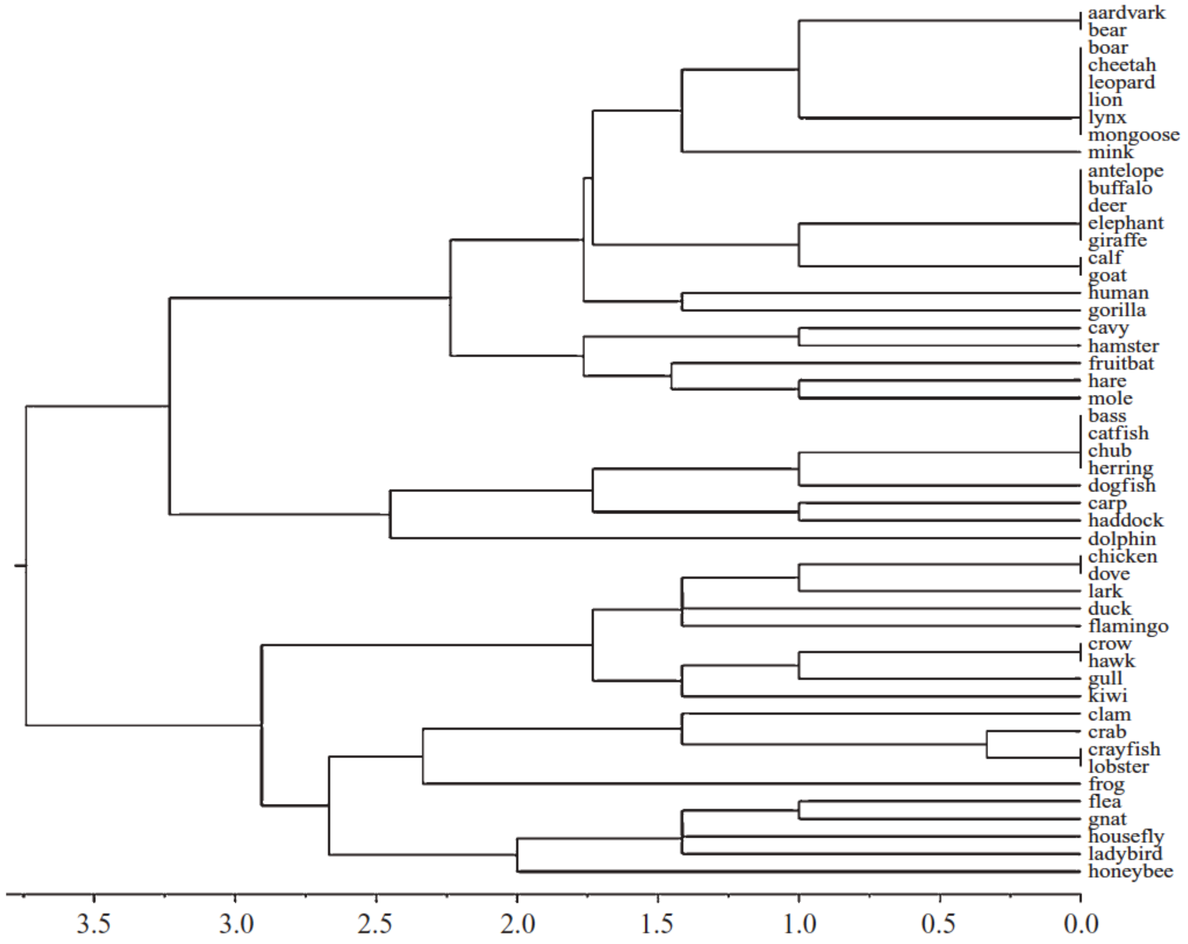
WITTEN, I. H.; FRANK, E.; HALL, M. A. ***Data Mining: Practical machine learning tools and techniques.*** Third Edition. Morgan Kaufmann, 2011. ISBN 978-0-12-374856-0.

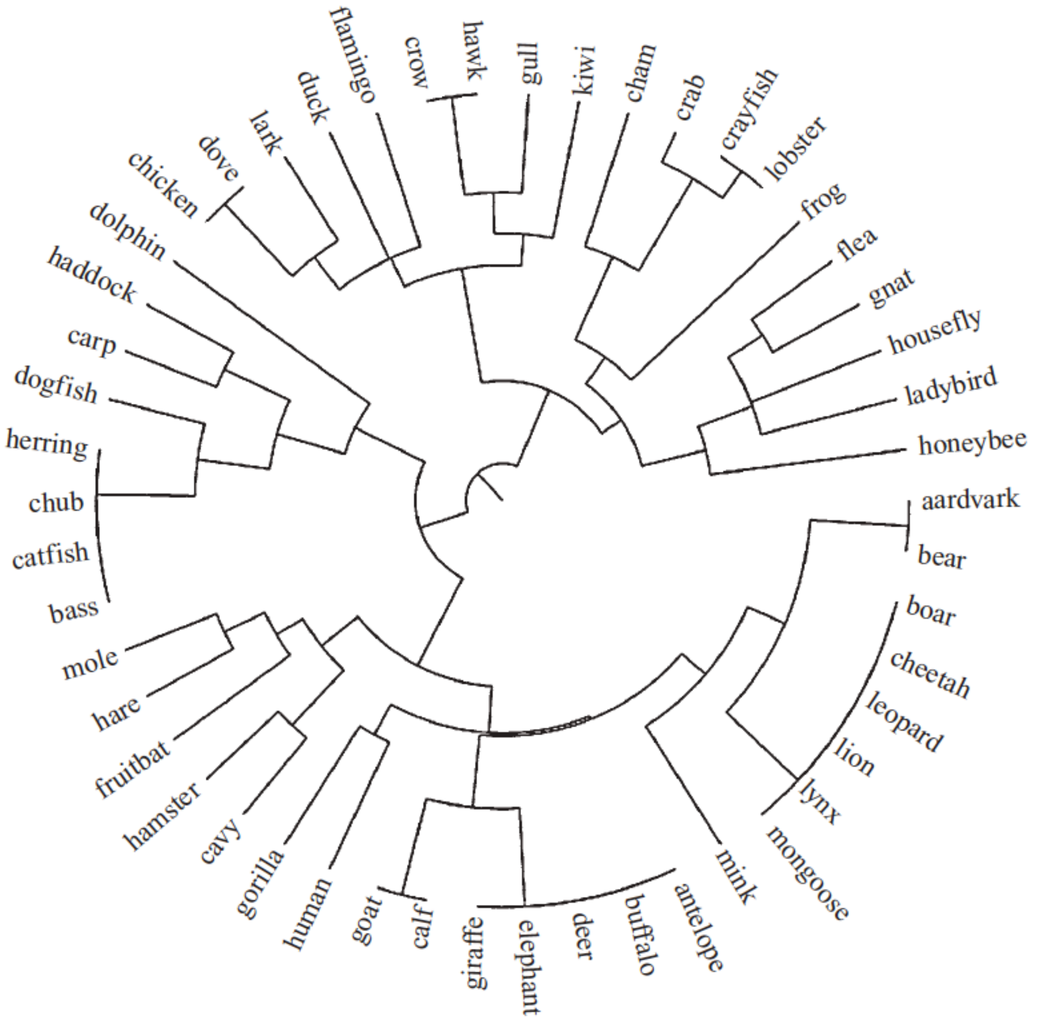
Dendogramas





Dendogramas





Agrupamiento jerárquico

- Una alternativa al método top-down para formar una estructura jerárquica de grupos es utilizar un **enfoque bottom-up**, que se llama la agrupación de aglomeración.
  - Esta idea fue propuesta hace muchos años y ha disfrutado recientemente un resurgimiento en popularidad.
  - El algoritmo básico es simple. Todo lo que se necesita es una **medida de distancia** (o una medida de similitud) entre los dos grupos.
  - Se empieza asignando a cada caso como un grupo en sí mismo; luego se encuentran los dos grupos más cercanos, se **combinan**, y sigue haciendo esto hasta que sólo un grupo quede.
  - El registro de *mergings* forma una estructura en forma de una **agrupación dendrográmica binaria** jerárquica.

Agrupamiento jerárquico

- Hay numerosas posibilidades para la medida de distancia.
- Una de ellas es la **distancia mínima** entre los clusters.
  - La distancia entre sus dos miembros más cercanos.
  - Esto produce lo que se llama el algoritmo de clustering un solo vínculo.
  - Dado que esta medida tiene en cuenta sólo los dos miembros más cercanos de un par de grupos, el procedimiento es sensible a los valores atípicos: La adición de una sola nueva instancia puede alterar radicalmente toda la estructura de la agrupación.
  - Además, si se define el diámetro de un grupo como la mayor distancia entre sus miembros, un solo vínculo de agrupación puede producir clusters con diámetros muy grandes.

Agrupamiento jerárquico

- Otra medida es la **distancia máxima** entre los grupos, en lugar del mínimo.
  - Se consideran dos grupos cercanos sólo si todas las instancias en su unión son relativamente similares, a veces llamado el método de vinculación exhaustividad.
  - Esta medida, que también es sensible a los valores atípicos, busca racimos compactos con diámetros pequeños
  - Sin embargo, algunos casos pueden terminar mucho más cerca de otros clusters de lo que son para el resto de su propio cluster.

Cobweb

- El algoritmo Cobweb fue desarrollado por investigadores en el área de Machine Learning en la década de 1980 para agrupar objetos en un conjunto de datos objeto-atributo.

- El algoritmo Cobweb genera un dendrograma de agrupamiento llamado árbol de clasificación que caracteriza a cada grupo.

Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb

- El algoritmo Cobweb construye un árbol de clasificación de forma incremental mediante la inserción de objetos en el árbol de clasificación uno por uno.
- Al insertar un objeto en el árbol de clasificación, el algoritmo Cobweb atraviesa el árbol de arriba hacia abajo a partir del nodo raíz.
- En cada nodo, el algoritmo considera 4 operaciones posibles y selecciona la que obtiene el valor más alto de la función utilidad de la categoría (CU):
  - Insert
  - Create
  - Merge
  - Split

Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb

- El algoritmo Cobweb funciona basándose en la llamada función de utilidad de la categoría (CU) que mide la calidad de la agrupación

- Si dividimos un conjunto de objetos en m grupos, entonces el CU de esta partición en particular es:

$$CU(C_1, C_2, \ldots, C_k) = \frac{\sum_l \Pr[C_l] \sum_i \sum_j \overbrace{\left(\Pr[a_i = v_{ij} \mid C_l]^2 - \Pr[a_i = v_{ij}]^2\right)}^{\text{Improvement in probability estimate because of instance cluster assigment}}}{k}$$

Cobweb: operaciones

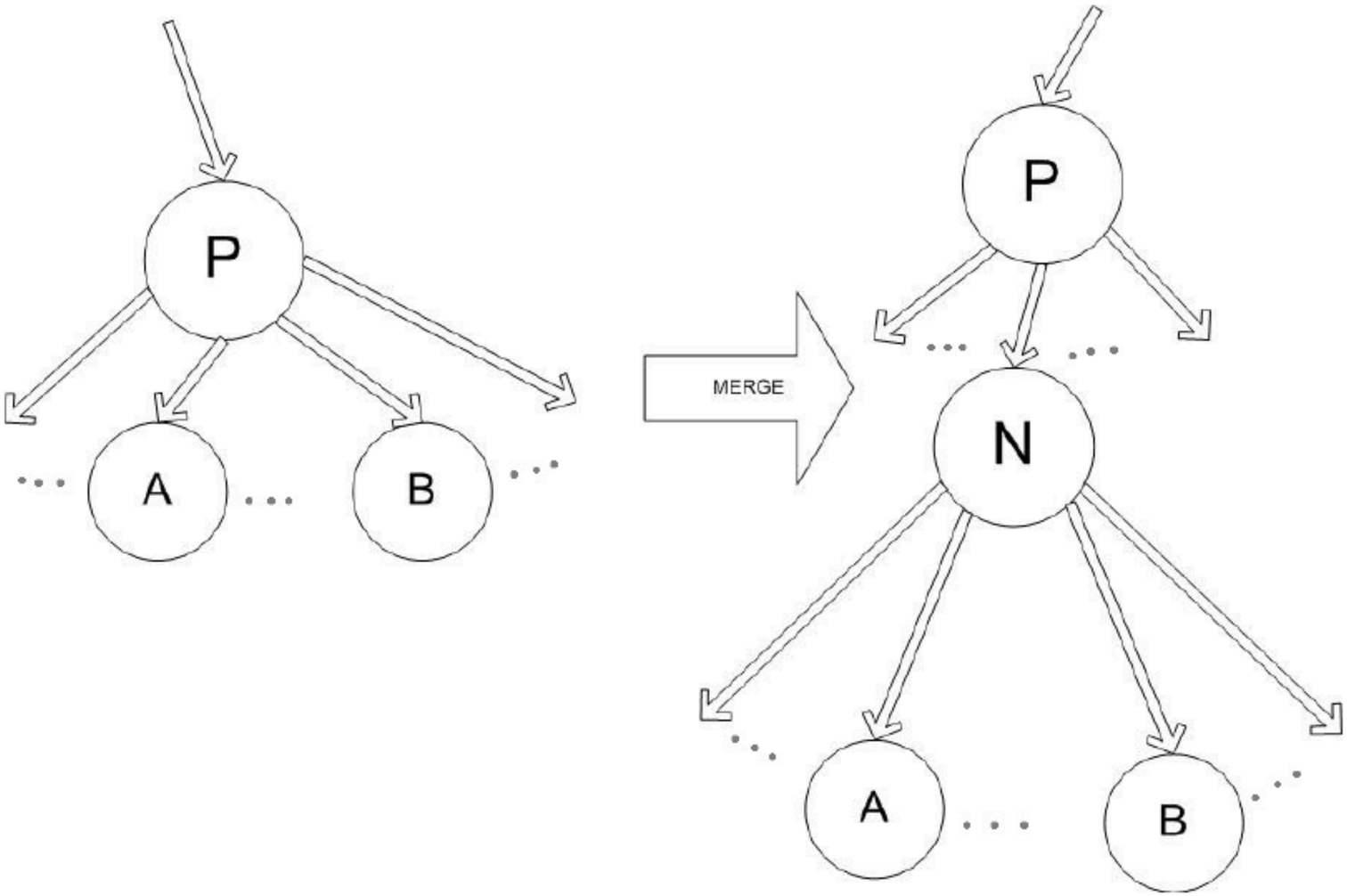
- Inserción significa que un nuevo objeto se insertará en uno de los nodos secundarios existentes.
  - El algoritmo evalúa el valor de la función CU de insertar el nuevo objeto en cada uno de los nodos secundarios existentes y selecciona el que tiene la puntuación más alta.
- El algoritmo también considera la creación de un nuevo nodo hijo específicamente para el nuevo objeto.

Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: operaciones

- El algoritmo Cobweb considera también la fusión (merging) de los dos nodos secundarios existentes con el más alto puntaje y con el segundo puntaje más alto.

Cobweb: operaciones



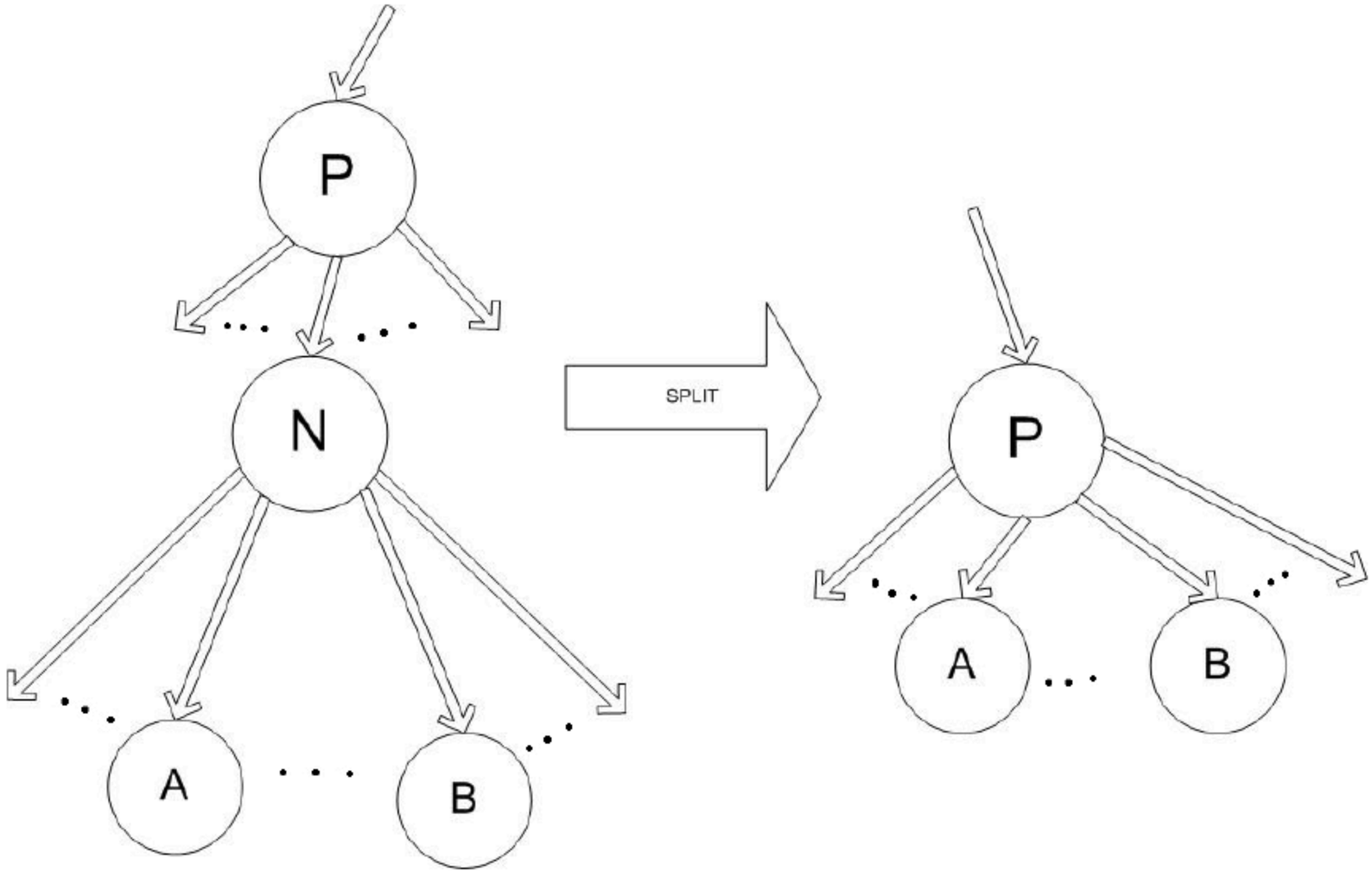
Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: operaciones

- El algoritmo Cobweb considera dividir (split) un nodo hijo existente con la puntuación más alta.

Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: operaciones



Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: algoritmo

- **Entrada**
  - El nodo actual N en la jerarquía de conceptos.
  - Una instancia I sin clasificar (atributo-valor).
- **Resultados**
  - Una jerarquía de conceptos que clasifica a la instancia.
- **Invocación**
  - Cobweb (nodo-raíz, I).
- **Variables**
  - C, P, Q, y R son nodos en la jerarquía.
  - U, V, W y X son los grupos (partición).

Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

```
Cobweb(N, I)
   Si N es un nodo terminal Entonces
      Crear-nuevo-nodo-terminal(N, I)
      Incorporar(N,I)
   Caso contrario
      Incorporar(N,I)
      Para cada hijo C del nodo N
         Calcular el puntaje de colocar I en C
      P = nodo con el puntaje más alto W
      Q = nodo con el segundo puntaje más alto
      X = puntaje de colocar I en un nuevo nodo R
      Y = puntaje por juntar P y Q en un nodo
      Z = puntaje por dividir P en sus hijos
      Si W es el mejor puntaje Entonces
         CobWeb (P, I)
      Caso contrario Si X es el mejor puntaje Entonces
         Colocar I como nuevo nodo
      Caso contrario Si Y es el mejor puntaje Entonces
         O = Merge(P, R, N)
         Cobweb (O, I)
      Caso contrario Si Z es el mejor puntaje Entonces
         Split(P, N)
         CobWeb (N, I)
```

Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

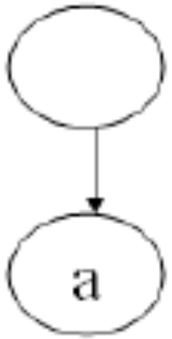
Cobweb: aplicación

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

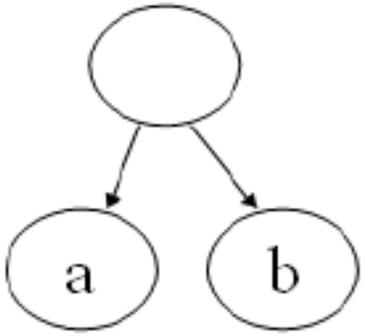
Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: aplicación

- Se inicia colocando la instancia en su propio cluster.



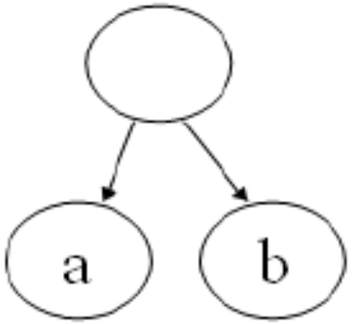
- Luego se añade la segunda instancia en su propio cluster.

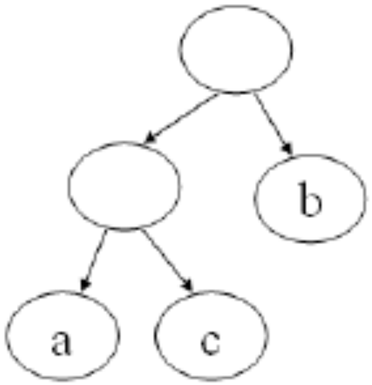


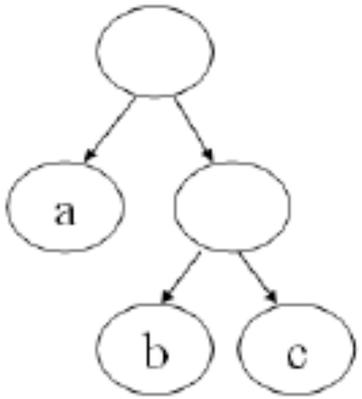
Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

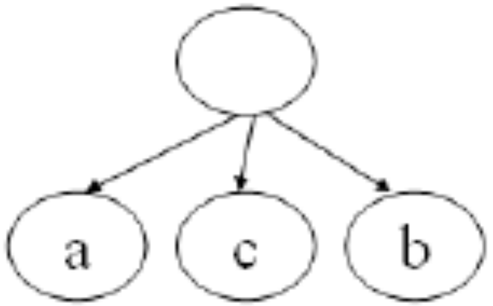
Cobweb: aplicación

- Agregando la tercera instancia.
  - Se evalúa la función utilidad de la categoría de:
    - Agregar la instancia en el primer cluster.
    - Agregar la instancia en el segundo cluster.
    - Agregar la instancia en su propio cluster.







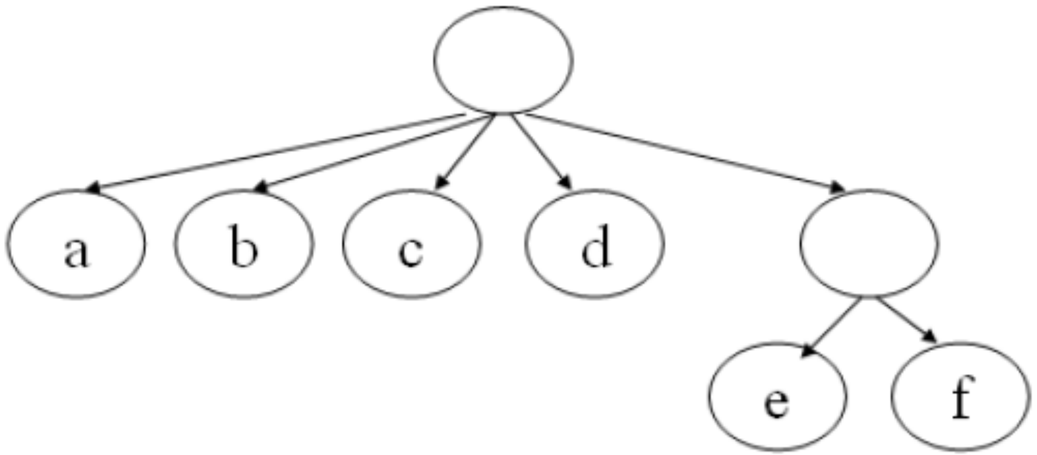


Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: aplicación

- Agregando la instancia f.

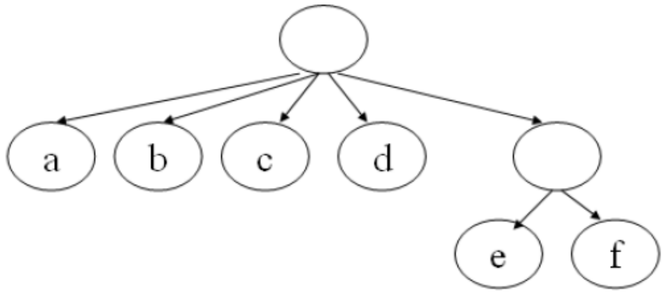
```
E) Rainy Cool Normal FALSE
F) Rainy Cool Normal TRUE
```



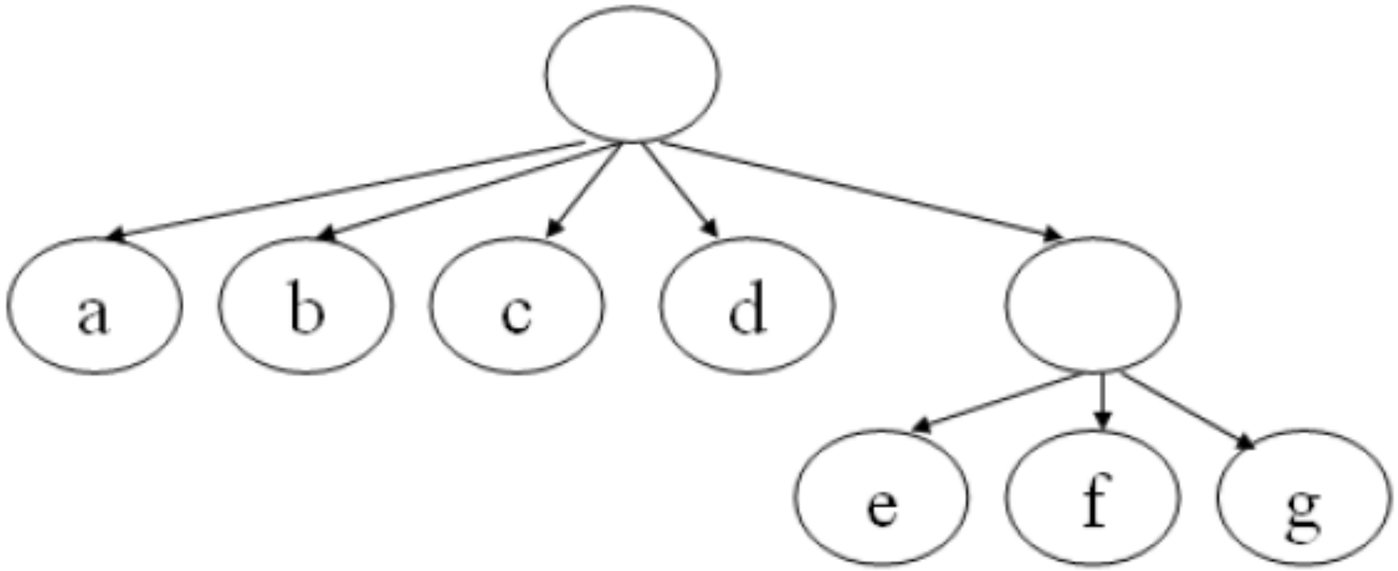
Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: aplicación

- Agregando la instancia g.



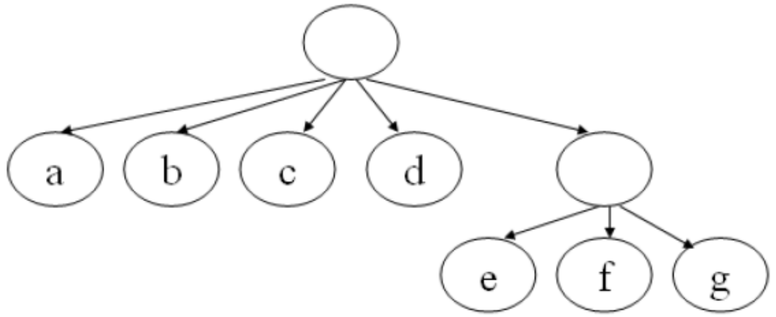

```
E) Rainy Cool Normal FALSE
F) Rainy Cool Normal TRUE
G) Overcast Cool Normal TRUE
```



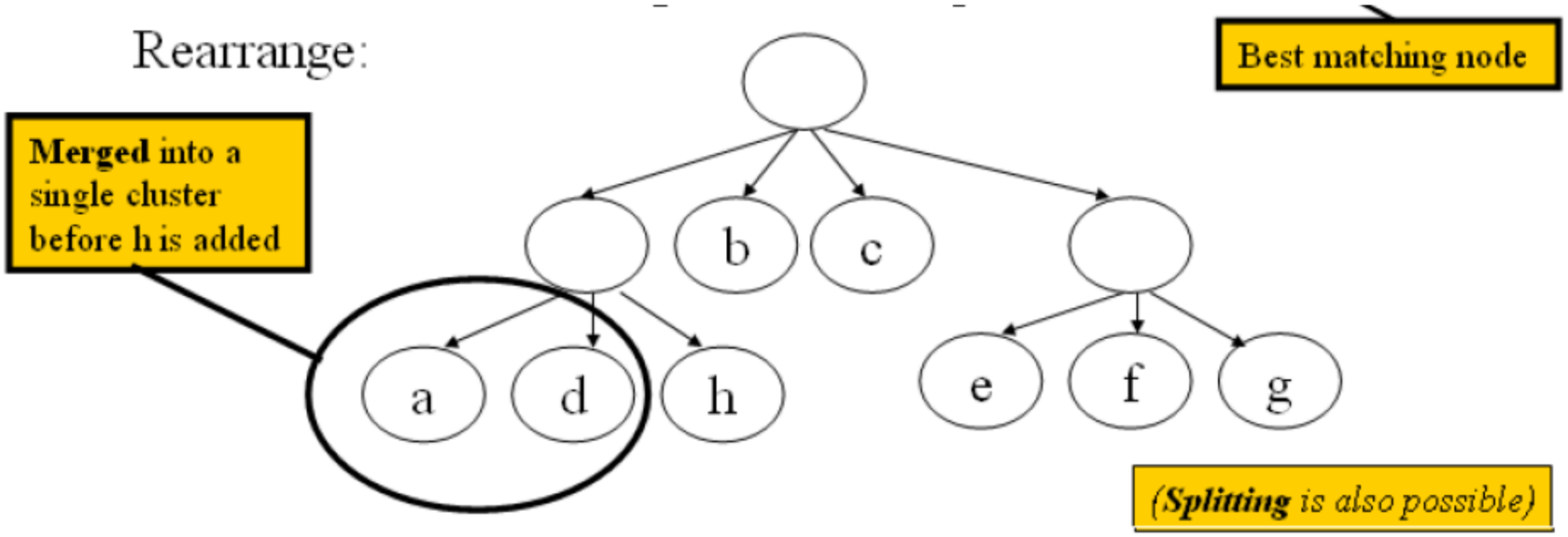
Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Cobweb: aplicación

- Agregando la instancia h.




```
A)  Sunny Hot High FALSE
D)  Rainy Mild High FALSE
H)  Sunny Mild High FALSE
```



Adaptado de http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab7-HirarchicalClustering.pdf

Competencias a adquirir en la sesión

- Al finalizar la sesión el alumno comprenderá el funcionamiento del **aprendizaje inductivo**.
- Al finalizar la sesión el alumno implementará **modelos algorítmos** usando algoritmos no supervisados.
- Al finalizar la sesión el alumno **entenderá** el algoritmo de **cobweb**.
- Al finalizar la sesión el alumno **aplicará** el algoritmo de **cobweb** para obtener modelos algorítmicos.

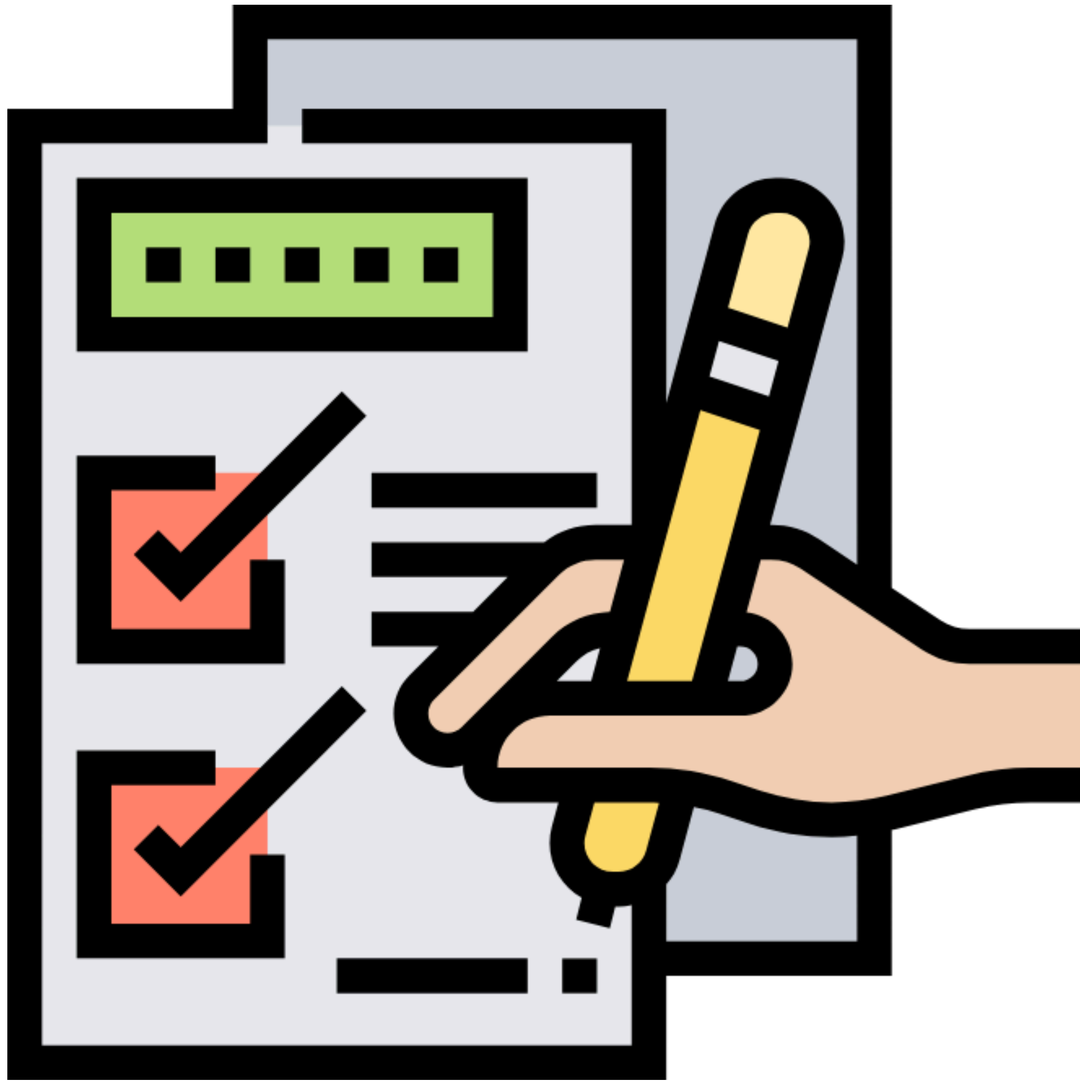


Imagen obtenida a través de la URL https://www.flaticon.com/free-icon/checklist_1467445