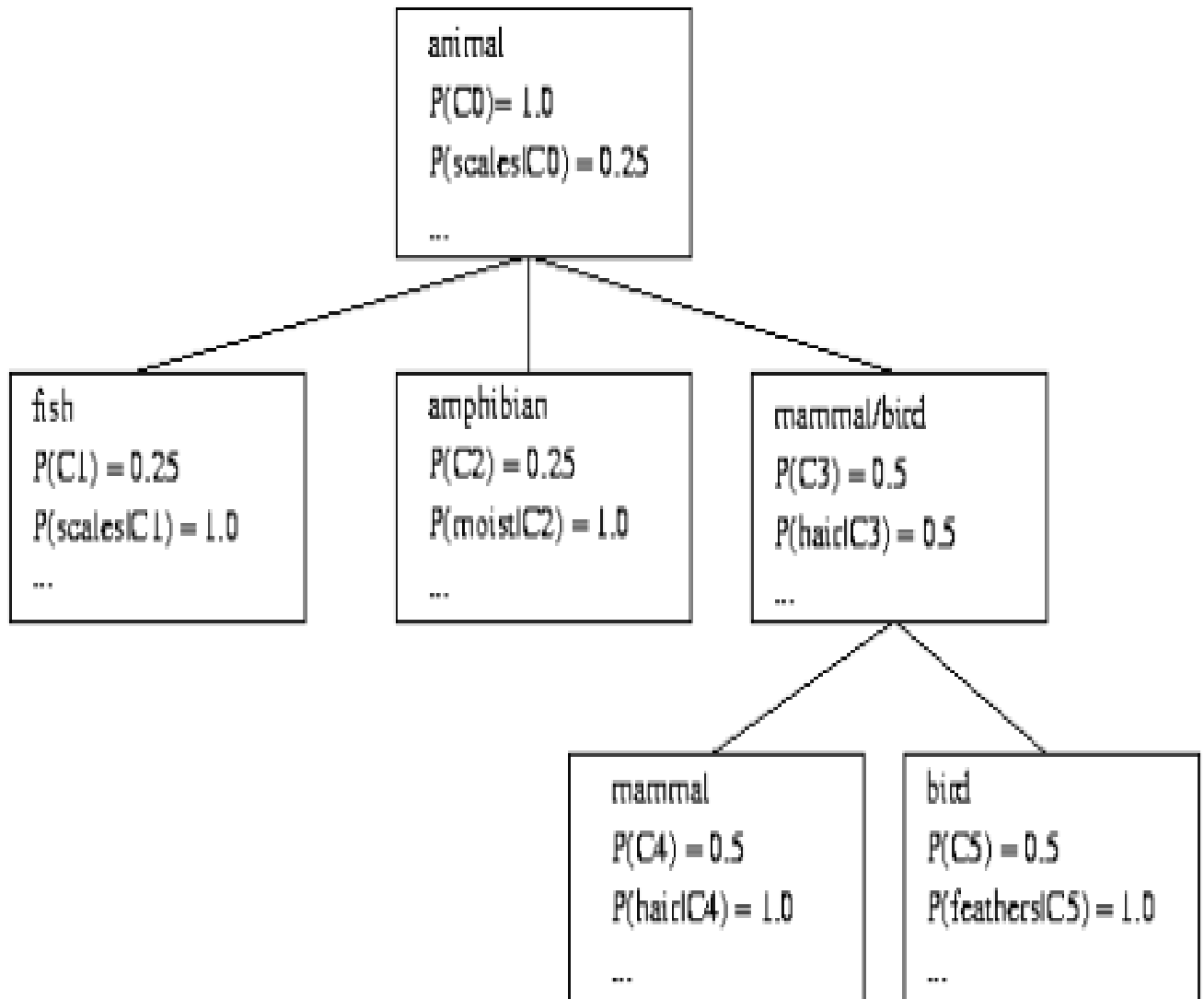# MODEL-BASED METHODS

- Attempt to optimize the fit between the given data and some mathematical model

- Based on the assumption: Data are generated by a mixture of underlying probability distribution

- In this method a model is hypothesize for each cluster and find the best fit of data to the given model

- This method also serves away of automatically determining number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

- Typical methods: EM, SOM, COBWEB

- EM — A popular iterative refinement algorithm

Starts with an initial estimate of the parameter vector

☐ Iteratively rescores the patterns against the mixture density produced by the parameter vector

☐ The rescored patterns are used to update the parameter updates

☐ Patterns belonging to the same cluster, if they are placed by their scores in a particular component

☐ Algorithm converges fast but may not be in global optima

☐ COBWEB

☐ A popular a simple method of incremental conceptual learning

☐ Creates a hierarchical clustering in the form of a classification tree

☐ Each node refers to a concept and contains a probabilistic description of that concept

animal

$P(C0)= 1.0$

$P(scales|C0) = 0.25$

...

fish

$P(C1) = 0.25$

$P(scales|C1) = 1.0$

...

amphibian

$P(C2) = 0.25$

$P(moist|C2) = 1.0$

...

mammal/bird

$P(C3) = 0.5$

$P(hair|C3) = 0.5$

...

mammal

$P(C4) = 0.5$

$P(hair|C4) = 1.0$

...

bird

$P(C5) = 0.5$

$P(feathers|C5) = 1.0$

...

SOM (Soft-Organizing feature Map)

- Competitive learning
- Involves a hierarchical architecture of several units (neurons)

- Neurons compete in a— winner- takes- all‖ fashion for the object currently being presented
- SOMs, also called topological ordered maps, or Kohonen Self-Organizing Feature Map
- (KSOMs)

MC4020 Data Mining and Data Warehousing Techniques

- It maps all the points in a high- dimensional source space into a 2 to 3- d target space, s.t the distance and proximity relationship (i.e., topology) are preserved as much as possible

- Similarity ok-means: cluster centers tend to lie in a low- dimensional fold in the feature space

- Clustering is performed by having several units competing for the current object

- The unit whose weight vector is closest to the current object wins

- The winner and its neighbors learn by having their weights adjusted

- SOMs are believed to resemble processing that can occur in the brain

- Useful for visualizing high-dimensional data in 2-or3-D space