K prototype

Just like K means where we allocate the record to the closest centroid here we allocate the record to the cluster which has the most similar looking reference point also known as prototype of the cluster also known as centroid of the cluster. More than similarity the algorithms try to find the dissimilarity between data points and try to group points with less dissimilarity into a cluster.

The dissimilarity measure for numeric attributes is the square Euclidean distance whereas the similarity measure on categorical attributes is the number of matching attributes between objects and cluster prototypes.

$D(x,p) = E(x,p) + \lambda C(x,p)$

where

x = Any datapoint,

y = Prototype of a cluster,

D(x,p) = Dissimilarity measure between x and y,

E(x,p) = Euclidean distance measure on the numeric attributes i.e. Euclidean distance between continuous attributes of x and y,

C(x,p) = Simple matching dissimilarity measure on the categorical attributes i.e. number of mismatched categorical attributes between x and y,

λ= weightage for categorical variable value. The weight is used to avoid favouring either type of attribute.

Steps

1. Number of clusters are chosen

2. It randomly initialize the cluster centers i.e. centroids (prototypes) of each cluster. The popular approaches are Huang and Cao approaches.

2. Upon initializing the initial prototypes (centroids) we will check each data point w.r.t each centroid and assign the data point to the cluster whose centroid dissimilarity is the least. After allocation, the centroid is updated by taking mode values of individual attributes for each data points in that cluster.

4. So after the initial allocation of records, one can realize that the initial centroids and the current centroid of a cluster would have changed. So with the current set of centroids, we will reallocate data points to a different cluster if the new prototype shows more similarities.

It halts creating and optimizing clusters when either:

- The centroids have stabilized - there is no change in their values because the clustering has been successful

- The defined number of iterations has been achieved.