ClusteringError Documentation

Francesco Turini

in corso

Contents

1 Binding energy based clustering

1.1 Clustering Problematics in physics studies

In physics studies, the mean of the measure is usually associated with its error. This additional information is not used in standard clustering algorithms like k-means, fuzzy C-means, or DBSCAN. In general, we want that more 'precise' is a measure more important is that measure in the clustering process. S

1.2 Basic Priciples

$$mass = \frac{\min(error)}{error}$$

$$Parallax \overset{[\frac{1}{\mathrm{errParallax}}]}{\rightarrow} Parallax$$

$$\phi_{ij} = \frac{mass_j}{r_{ij}}$$

1.2.1 Distances normalization

If we want use different feature of a dataset, we must think how to define a distance between the different data points. For example if a data point is defined by two variables $X\,[a,b]$ and $Y\,[c,d]$, if we have $a,b \gg c,d$ whene we go to calculate the distance between two points the feature Y dosen't affect much the distance, so is like we don't use the information inside Y.

For avoiding that is usefull to normailze the data before to calculate the distances. A good practice is to have a zero-mean distribution with a unitary standard deviation even if we loss information about the dispersion of the data distribution. At the end the normalization choosen is:

$$X_i = \frac{X_i - \bar{X}_i}{\sigma_{X_i}} \tag{1}$$

$$\sigma_{X_i} = \sqrt{E[(X_i - \bar{X}_i)^2]} = \sqrt{\frac{1}{len(X_i)}\sum_j (X_{i,j} - \bar{X}_i)^2} \tag{2}$$

where the index i mean the feature.