数据标准化处理

标准化是将属性域里面的数据等比例缩放，使得处理后的值落入一个小的特定区间。标准化主要有以下几点好处：

（1）、可以将有单位的属性变成无单位的，这样就可以均等的对待每一个属性。比如对吞吐量量化之后的值进行标准化，不仅可以去掉单位，而且使得不同的属性值可以一起参加计算。

（2）、很好地解决了大数吃小数问题。

（3）、可以加快数据的运算速度。当将标准化之后的数据放到Matlib中计算，可以很明显的加快数据的运算处理速度。

标准化的方法有很多种，比较常用的区间度量属性标准化有两种：范围标准化以及Z-score标准化。

范围标准化

范围标准化也叫做离差标准化，它的核心思想是将同一个属性值域集合里面的不同属性值转变到范围为0到1之间的数。假设属性值域集合为X(S)；那么对集合X(S)中的值进行标准化公式为：

$$x_i' = \frac{x_i - min(X)}{max(X) - min(X)}$$

数据标准化

其中，x_i表示标准化之前的属性量化值，x_i'为标准化之后的属性量化值。min(X)表示集合X(S)中数值最小的值，同理max(X)表示集合X(S)中数值最大的值。max(X) - min(X)表示属性值域变化的范围。范围标准化之后，集合X(S)中的每一个值变
化范围将
在0到1之间。范围
标准化的缺陷是当有新数据加入到集
合里面时，可能会导致max(X)和min(X
)的变化，这样需要重新标准化计算；但范围标准化的优点是可以把带有量纲的属性值标准化为无量纲的值。

Z-score标准化

Z-score标准化也被称为标准差标准化，这是因为Z-score标准化公式中利用到属性的标准差和平均值来计算的。Z-score标准化可以判断出集合中的数值是从x坐标轴的正方向还是反方向远离属性的平均值。和范围标准化方法不同的是，Z-
score标准化之后还是有量纲的，单位量是属性的标准差。Z-score标准化的计算公式如下所示：

$$x_i' = \frac{x_i - \mu_X}{\sigma_X}$$

其中，$\mu_X = \frac{1}{n}\sum_{i=1}^{n} x_i$，$\sigma_X = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_X)^2}{n-1}}$

数据标准化

式中x_i表示标准化之前的值，x_i'为标准化之后的值；μ_x代表集合X(S)中数据的平均值；σ_x代表了集合X(S)中数据的标准差。Z-score标准化后的数据经过处理后服从标准正态分布