在不知道数据标签情况下的一种无监督指标

假设我们将数据集Z（z1,...,zN）=Z（x1,...xn,y1,...,ym）(长度N=n+m)通过分类方法分为子集X（x1,...xn）（长度为n）和子集Y（y1,...,ym）（长度为m）。下面介绍一种无监督方法来判断分类效果的好坏。该方法主要从均值方面考虑分类后的两个子集的分布是否不一致。也即分类后均值是否不同。

记X，Y，Z的均值为x-，y-，z-。记S12 =∑i=1,..,n (xi - x-)2， S22 =∑i=1,..,m (yi - y-)2，S32 =n (x- - z-)2 +m(y- - z-)2，S2 =∑i=1,..,N (zi - z-)2。直接展开可验证S2 = S12 + S22 + S32。S12、S22为组内距离，S32为组间距离，S2为总距离。如果分类后的两个子集的分布一致，则S32为0，此时分类效果差。反之分类越好，S32越大。我们将S32/S2作为分类的效果指标。该比值可用于粗略比较两种分类方法的好坏，但该比值并不是分类准确率。