Mahalanobis 距离

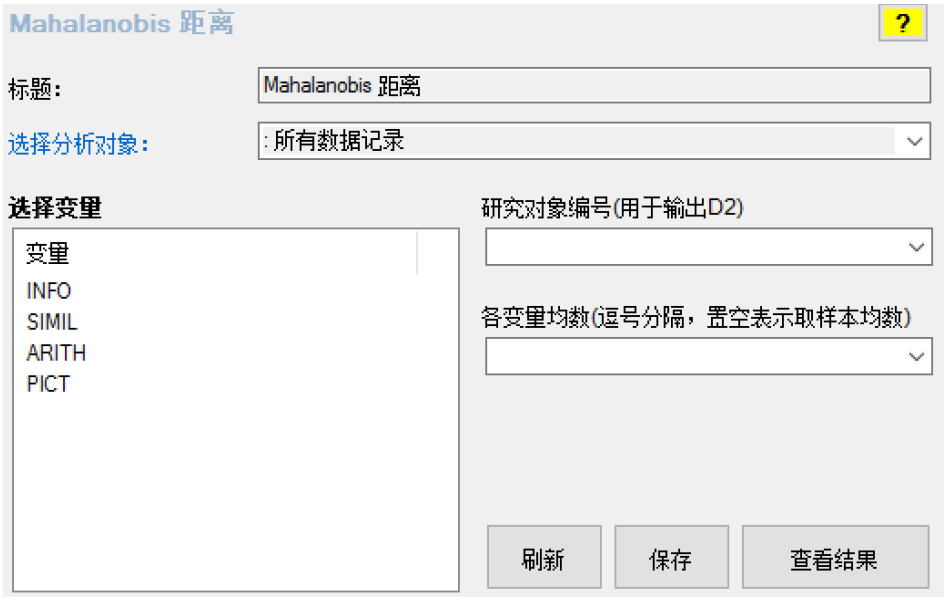
在数据分析和数据挖掘的过程中，经常需要知道个体间差异的大小，进而评价个体的相似性和类别。如我们要比较 X 个体和 Y 个体间的差异，它们都包含了 N 个维的特征，即 X=（x1，x2， x3， … xn），Y=（y1， y2， y3， … yn）。Mahalanobis 距离是由印度统计学家马哈拉诺比斯（P. C. Mahalanobis）提出的，是一种采用协方差来计算多维空间中两点之间距离的方法，它是一种有效的计算两个未知样本集的相似度的方法。马氏距离在多元正态分布资料分析时应用广泛，虽然该方法并不要求数据呈正态分布。Mahalanobis 考虑了各变量之间的相关性，并且与各变量的单位无关。（例如：一条关于身高的信息会带来一条关于体重的信息，因为两者是有关联的并且是尺度无关的）。使用 Mahalanobis 距离的好处是可以排除变量之间的相关性的干扰，由标准化数据和中心化数据（即原始数据与均值之差）计算出的二点之间的 Mahalanobis 距离相同。

本模块中 Mahalanobis 距离采用 D2 指标，用于计算某个观测值与样本均值之间的距离。

D2= (x - u)’ -1(x - u)

利用 Mahalanobis 距离 D 可以方便地检测出多维空间中的极端值（观测值与样本均值间有显著性差异）。

例：练习项目 wais 计算 Mahalanobis 距离，输入界面如下：



输出结果：

总记录数： 40

排除含缺失值的记录： 0

	INFO	SIMIL	ARITH	PICT
比较均数	11.25	8.425	10.575	7.15
样本均数	11.25	8.425	10.575	7.15

Mahalanobis 距离(D2)最大的 10 条记录：

_TMPID	D2(Mahalanobis 距离)	P 值
32	11.6978883548012	0.0197451048965988
5	9.76017786328915	0.0446671455977531
3	9.43601061814374	0.0510788231880603
8	9.19175823888347	0.0564811308359144
33	7.50804752909664	0.111354953905694
36	7.18455164211063	0.126451037853392
28	6.10250116496361	0.191623079651231
39	5.5963962857185	0.231385214639087
22	5.55422582721436	0.235003945580933
2	5.09889662189923	0.277299778563599

下列记录的 Mahalanobis 距离(D2)的 p 值<0.05：

_TMPID	D2(Mahalanobis 距离)	P 值
5	9.76017786328915	0.0446671455977531
32	11.6978883548012	0.0197451048965988