

DOI:10.19454/j.cnki.cn15-1170/c.2021.06.005

多元统计方法是否需要变量进行加权

——以判别分析和聚类分析为例

○ 文 / 李子宁

文章以判别分析和聚类分析为例，在理论上证明了对变量加权是否会对结果产生影响，并进行了实证分析。研究表明，是否对变量加权不影响判别分析结果，但影响聚类分析结果。这一结论可进一步拓展，即凡是以马氏距离为基础的方法不需要对变量进行加权，而以欧氏距离为基础的方法如果对变量进行加权可以提高分析结果的准确度。

机

器学习是一门新兴的交叉学科，它既包括一些传统的多元统计方法，如聚类分析、判别分析、逻辑回归、因子分析等，也包括一些人工智能方法，如K近邻法、决策树、人工神经网络、支持向量机等。在这些方法中，也许K近邻法是最简单的方法，它的基本思想是以K个最近邻居在因变量取值的平均数作为新样品的预测值。它又派生出基于变量重要性的加权K近邻和基于观测重要性的加权K近邻。由于其它统计方法都不涉及邻居，因此基于观测重要性的加权方法不具有外推性。那么基于变量重要性的加权方法是否具有外推性呢？或者说，我们常用的判别分析、聚类分析等需不需要对变量进行加权呢？文章将对此问题进行理论和实证分析。

一、基于变量重要性加权的基本原理：以K近邻法为例

（一）变量重要性的确定方法

变量的重要性可以从三个方面进行考察，一是从变量本身考察，二是从解释变量与被预测变量的相关性角度考察，三是从预测误差角度考察^[1]。

从变量自身来考察，变异程度最大的变量重要性更强，如果一个变量是常数，没有什么变异，则这个变量对预测是没有意义的。对数值型变量来说，衡量变异性的常用指标是方差、标准差和变异系数，由于方差和标准差受计量单位的影响，在衡量变量重要性时并不适用，通常采用变异系数，即变异系数越大的变量越重要。对于类别变量，如果各个类别值的取值

比例相当，则这个变量越重要；如果某个类别的取值比例越大，则这个变量越不重要。以二分类变量为例，如果两个类别的取值比例均为0.5，此时这个类别变量的方差取最大值0.25；而如果一个类别所占比例为0.9，另一个类别所占比例为0.1，此时这个类别变量的方差仅为0.09。

从解释变量与被预测变量的相关性角度来考察，又可以分成三种情况。第一种情况是解释变量与被预测变量均为类别变量。衡量类别变量间相关与否的统计量为卡方统计量，卡方统计量越大，类别变量间的相关程度就越大，因此卡方越大的变量或p值越小的变量越重要。第二种情况是解释变量与被解释变量均为连续变量。连续变量相关与否的统计量为相关系数，相关系数越大，变量间的相关性越强；当然前提是相关系数必须是显著的，这可以通过t统计量进行检验。第三种情况是解释变量和被预测变量分属不同类别，具体包括两类：解释变量是类别变量，被预测变量是连续变量；解释变量是连续变量，被预测变量是类别变量。无论是两种情况中的哪一种，均采用方差分析的方法，即计算F统计量，F统计量越大，表明变量之间相关性越强。

从预测误差角度来考察，通常与建模策略有关。建模策略有两种，一是“从一般到具体”建模策略，二是“从具体到一般”建模策略。若采用“从一般到具体”建模策略，首先将全部变量加入模型，然后分别去掉一个解释变量，建立K个K-1元模型，在这K个K-1元模型中，哪个模型的预测误差最大，说明该模型所不包含的那个变量重要性越大。若采用“从具

体到一般”建模策略,则可直接比较 K 个一元模型,哪个模型的拟合程度越好(即误差越小),即说明哪个变量的重要性越大。一般认为,“从一般到具体”建模策略更好,因为“从具体到一般”建模策略可能会造成遗漏变量问题。

(二) 变量权重的确定方法

根据变量重要性的确定方法,令第 i 个解释变量的权重为 w_i ,它是解释变量重要性的函数,可定义为:

$$w_i = \frac{FI_i}{\sum FI_i}$$

其中 FI_i 为解释变量重要性,从机器学习角度又被称为特征重要性,它以输入变量对预测误差的影响定义。假定有 K 个输入变量, x_1, x_2, \dots, x_k , 剔除第 i 个变量,计算输入变量为 $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_k$ 下, K 近邻法的错判概率,记作 e_i 。若第 i 个变量对预测有重要作用,剔除该变量后的预测误差将比较大。因此第 i 个变量的重要性定义为: $FI_i = e_i + \frac{1}{k}$ 。因此不论从哪个角度来考察,变量越重要,在计算距离时其权重越大。

由于 K 近邻法采用欧氏距离测度近邻观测,则加权的欧氏距离为:

$$EUCLID(x, y) = \sqrt{\sum w_i (x_i - y_i)^2}$$

(三) 使用 K 近邻法进行预测

对于二分类预测问题,如果有超过半数的近邻类别值为 1,则预测值为 1 类,否则预测值为 0 类。对于多分类预测问题,预测值为众数。对于回归预测问题,预测值是 K 个近邻在被预测变量上的平均值。

二、判别分析是否需要变量进行加权

判别分析是指在已知研究对象分成若干组的情况下,判断新的样品应归属的组别。在判别分析中,最直观的判别方法就是距离判别,即计算新样品到各组的距离,新样品距离哪组最近,就被判为哪一组。

(一) 两组距离判别

设组 π_1 和 π_2 的均值分别为 μ_1 和 μ_2 , 协方差矩阵分别为 Σ_1 和 Σ_2 , x 是一个新样品,现判断它来自哪一组。

若不对变量进行加权,计算 x 到两个组的距离 $d^2(x, \pi_1)$ 和 $d^2(x, \pi_2)$,并按如下的判别规则进行判断^[1]:

$$\begin{cases} x \in \pi_1 & \text{若 } d^2(x, \pi_1) \leq d^2(x, \pi_2) \\ x \in \pi_2 & \text{若 } d^2(x, \pi_1) > d^2(x, \pi_2) \end{cases} \quad (1)$$

1. $\Sigma_1 = \Sigma_2 = \Sigma$ 时的判别。若对变量进行加权,设 w_i 为第 i 个判别变量的权重,则加权后的判别向量为 $x' = wx$, 均值向量为 $w\mu$, 方差协方差矩阵为 $w \Sigma w'$ 。

经过加权的平方马氏距离为:

$$\begin{aligned} d^2(x^*, \pi_1) &= [w(x - \mu_1)]' (w \Sigma w')^{-1} [w(x - \mu_1)] \\ d^2(x^*, \pi_2) &= [w(x - \mu_2)]' (w \Sigma w')^{-1} [w(x - \mu_2)] \end{aligned}$$

其中 w 为对角矩阵:

$$w = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \dots & \\ & & & w_p \end{bmatrix}$$

判别规则为:

$$\begin{aligned} x \in \pi_1 & \quad \text{若 } d^2(x^*, \pi_1) \leq d^2(x^*, \pi_2) \\ x \in \pi_2 & \quad \text{若 } d^2(x^*, \pi_1) > d^2(x^*, \pi_2) \end{aligned} \quad (2)$$

将加权的平方马氏距离展开:

$$\begin{aligned} d^2(x^*, \pi_1) &= (x - \mu_1)' w' (w')^{-1} \Sigma^{-1} w^{-1} w (x - \mu_1) \\ &= (x - \mu_1)' \Sigma^{-1} (x - \mu_1) \\ &= d(x, \pi_1) \\ d^2(x^*, \pi_2) &= (x - \mu_2)' w' (w')^{-1} \Sigma^{-1} w^{-1} w (x - \mu_2) \\ &= (x - \mu_2)' \Sigma^{-1} (x - \mu_2) \\ &= d(x, \pi_2) \end{aligned}$$

由于 $d(x^*, \pi_1) = d(x, \pi_1)$, $d(x^*, \pi_2) = d(x, \pi_2)$ 。所以在两组距离判别且假定方差阵相等时,对变量加权并不影响判别分析的结果。

2. $\Sigma_1 \neq \Sigma_2$ 时的判别。若对变量进行加权,则

$d^2(x^*, \pi_1)$ 和 $d^2(x^*, \pi_2)$ 的计算公式为:

$$\begin{aligned} d^2(x^*, \pi_1) &= [w(x - \mu_1)]' (w \Sigma_1 w')^{-1} [w(x - \mu_1)] \\ d^2(x^*, \pi_2) &= [w(x - \mu_2)]' (w \Sigma_2 w')^{-1} [w(x - \mu_2)] \end{aligned}$$

$d^2(x^*, \pi_1)$ 和 $d^2(x^*, \pi_2)$ 可以进一步整理为:

$$\begin{aligned} d^2(x^*, \pi_1) &= (x - \mu_1)' w' (w')^{-1} \Sigma_1^{-1} w^{-1} w (x - \mu_1) \\ &= (x - \mu_1)' \Sigma_1^{-1} (x - \mu_1) \\ &= d(x, \pi_1) \\ d^2(x^*, \pi_2) &= (x - \mu_2)' w' (w')^{-1} \Sigma_2^{-1} w^{-1} w (x - \mu_2) \\ &= (x - \mu_2)' \Sigma_2^{-1} (x - \mu_2) \\ &= d(x, \pi_2) \end{aligned}$$

因此在两组距离判别且方差阵不相等时,对变量加权也不影响判别分析的结果。

(二) 多组距离判别

设有 k 个组 $\pi_1, \pi_2, \dots, \pi_k$, 它们的均值分别为 $\mu_1, \mu_2, \dots, \mu_k$, 协方差矩阵分别是 $\Sigma_1, \Sigma_2, \dots, \Sigma_k$, x 到总体 π_i 的加权平方马氏距离为:

$$d^2(x^*, \pi_i) = [w(x - \mu_i)]' (w \Sigma_i w')^{-1} [w(x - \mu_i)]$$

判别规则为:

$$x \in \pi_i, \quad \text{若 } d^2(x^*, \pi_i) = \min_{1 \leq i \leq k} d^2(x^*, \pi_i)$$

由于 $d^2(x^*, \pi_i) = d^2(x, \pi_i)$,所以在多组距离判别下,对变量加权与不影响判别结果。

三、聚类分析是否需要变量进行加权

聚类分析是一种无监督学习方法,没有目标变量,因此聚类分析中一般不采用马氏距离,而采用欧氏距离。但欧氏距离与各变量的量纲有关,没有考虑变量间的相关性,也没有考虑各变量方差的不同^[3]。因此对变量是否加权会影响聚类结果。

当不对变量进行加权时,两个样品之间的平方欧氏距离为:

$$d^2(x, y) = (x - y)'(x - y)$$

当对变量进行加权时,两个样品之间的平方欧氏距离为:

$$d^2(x^*, y^*) = [w(x - y)]' [w(x - y)]$$

通常 $d^2(x, y) \neq d^2(x^*, y^*)$, 因此两种情况下的聚类结果一般不会相同。

若对数据进行标准化,令 $x^s = \sum^{-\frac{1}{2}}(x - \mu)$, $y^s = \sum^{-\frac{1}{2}}(y - \mu)$, 显然有 $E(x^s) = E(y^s) = 0$, $V(x^s) = V(y^s) = I$, 则

$$\begin{aligned} d^2(x^s, y^s) &= [\sum^{-\frac{1}{2}}(x - \mu) - \sum^{-\frac{1}{2}}(y - \mu)]' [\sum^{-\frac{1}{2}}(x - \mu) - \sum^{-\frac{1}{2}}(y - \mu)] \\ &= (x - y)' \sum^{-1} (x - y) \\ &= (x - y)' \sum^{-1} (x - y) = d^2(x, y) \end{aligned}$$

因此,变量标准化之后的平方欧氏距离等价于标准化前的平方马氏距离。所以在进行聚类分析时,如果选择对变量进行标准化,是否对变量进行加权对聚类分析结果没有影响。

四、实证分析

(一) 对加权和不加权情况下判别分析的验证

文章以费希尔判别分析的经典例子来验证对变量加权与否的影响。费希尔于 1936 年发表的鸢尾花数据包括 3 种鸢尾花: 刚毛鸢尾花、变色鸢尾花和弗吉尼亚鸢尾花, 每种各抽取一个容量为 50 的样本, 测量了花萼长、花萼宽、花瓣长、花瓣宽 4 个变量。

无论采用何种判别方法, 都可能会产生误判。误判比例的计算通常有四种方法, 一是直接用样本计算判别函数, 同时计算误判比例, 这种方法给出的误判比例通常较低; 二是旁置法, 即拿出样本的一部分 (通常为 70%) 作为训练样本集构造判别函数, 剩余的部分作为测试样本集计算误判比例; 三是十折交叉验证法, 即将样本分成十部分, 每次取其中的九部分作为训练样本集构造判别函数, 剩余的一部分作为测试样本集计算误判比例, 十折交叉验证要构造十个判别函数; 四是刀切法, 即每次拿出一个观测作为测试样本, 其余的观测作为训练样本集构造判别函数。刀切法避免了样本数据在构造判别函数的同时又被用来对该判别函数进行评价, 也几乎避免了构造判别函数时样本信息的损失。

文章采用第一种和第四种计算误判比例的方法。其中表 1 为未对变量进行加权的距离判别结果, 表 2 为按预测误差加权的距离判别结果。



表 1 未对变量加权的距离判别结果

组别		预测类别			Total
		1	2	3	
原始类别	频数	1	50	0	50
		2	0	48	50
		3	0	1	50
刀切法验证	频数	1	50	0	50
		2	0	48	50
		3	0	1	50

表 2 对变量加权的距离判别结果

组别		预测类别			Total
		1	2	3	
原始类别	频数	1	50	0	50
		2	0	48	50
		3	0	1	50
刀切法验证	频数	1	50	0	50
		2	0	48	50
		3	0	1	50

表 1 和表 2 中, 无论直接采用判别函数验证, 还是采用刀切法验证, 是否对变量进行加权的结果完全相同。

(二) 对加权和不加权情况下聚类分析的验证

我们仍然使用费希尔的数据, 其中编号 1-50 属刚毛鸢尾花, 编号 51-100 属变色鸢尾花, 编号 101-150 属弗吉尼亚鸢尾花。聚类变量为花萼长、花萼宽、花瓣长、花瓣宽 4 个变量, 聚类方法采用组间连接法, 聚类数目为 3 类。当未对变量进行加权时, 编号 1-50 仍被分到第一组, 编号 51-100 仍被分到第 2 组, 但编号 100-150 中只有 110、112、118、120、122、127、130、131、135、138、140、144 被分到第三组, 其余 38 个被错分到了第二组。当对变量进行加权时, 前 50 个观测仍被分到第一组, 编号 51-99 被分到第二组, 但编号 100 被分到了第三组; 编号 101-150 中只有 14 个被错误分到了第二组。因此对变量进行加权的聚类分析, 其聚类效果好于不对变量进行加权的聚类分析。另外, 在变量加权和不加权两种情况下, 如果在聚类分析时选择对变量进行标准化, 则结果完全相同。

五、结论与拓展

从理论和实证分析来看, 凡是采用马氏距离的方法, 都不需要对变量进行加权。凡是采用欧氏距离的方法, 如果不对变量进行标准化, 则是否加权影响分析结果; 若对变量进行标准化, 欧氏距离等同于马氏距离, 是否加权对分析结果没影响。

这一结论可以进一步拓展。比如典型判别, 实质是二阶段判别, 第一阶段降维, 第二阶段采用降维后的主成分进行距离判别。因此典型判别本质上仍是

距离判别, 由于距离判别采用马氏距离, 是否对变量进行加权并不影响典型判别的结果。对于 K 近邻法, 如果采用马氏距离, 则不需要对变量进行加权, 也就没有所谓的基于变量加权的 K 近邻法; 但目前统计软件都是基于欧氏距离或街区距离, 且默认对变量进行标准化, 此时对变量是否加权不影响结果; 如果不对变量进行标准化, 则基于变量加权的 K 近邻法和普通的 K 近邻法在分析结果上是有差异的。

对于因子分析和主成分分析, 其基本原理是对方差矩阵或相关矩阵进行分解。统计软件一般默认基于相关矩阵进行分析^[4], 此时是否对变量进行加权不影响结果; 但若基于协方差矩阵进行分析, 是否对变量加权会影响分析结果。

基金项目: 河北省高等教育教学改革与研究项目“经济统计专业推进课程思政的探索与实践(批准号: 2020GJJG590)”阶段性成果。

作者单位: 河北工业大学理学院

参考文献

- [1] 薛薇. R 语言数据挖掘方法及应用 [M]. 电子工业出版社, 2016.
- [2] 王学民. 应用多元分析 (第五版) [M]. 上海财经大学出版社, 2017.
- [3] 高惠璇. 应用多元统计分析 [M]. 北京大学出版社, 2005.
- [4] 李国柱. SPSS 统计分析基础 [M]. 国家开放大学出版社, 2018.