数据归约中基于因子分析的属性选择方法

刘云霞[1],贺晋兵[2]
(1.厦门大学 经济学院计划统计系,福建 厦门 361005;2.大公国际资信评估有限公司,北京 100016)

摘 要:属性子集的选择是数据归约的重要内容。文章提出了一种基于因子分析的无监督属性选择的方法。通过该方法选出的属性子集能够最好地覆盖数据的自然分类。在统计模拟中,这种方法也得到了很好的效果。

关键词:数据归约;因子分析;属性子集选择
中图分类号:C81 文献标识码:A 文章编号:1002-6487(2009)07-0036-02

属性子集的选择是数据归约的重要内容。目前,数据挖掘中各种属性选择算法的研究多是从粗糙集理论、仿生学、机器学习以及模式识别等角度进行的。这些方法在选择过程中几乎都涉及到分类,多是在有监督学习情形下进行的属性子集选择,而对于无监督学习情形下属性选择的研究则比较少。鉴于此,本文拟依据"无监督属性选择中属性子集最优的标准是该子集能够最好地覆盖数据的自然分类"[1]的原则,提出一种基于因子分析的属性选择方法来进行无监督属性子集的选择。

1 逐步选择方法的不足

在数据归约中较为常用的方法包括逐步向前选择、逐步向后删除以及向前选择和向后删除相结合的方法。这些方法虽然都是次优搜索的启发式方法中比较常用的技术,但也存在着一些不足。

逐步向前选择方法是一种自下而上的搜索方法。它由空属性集开始,依次从未入选的属性中选择一个属性,使它与已入选的属性组合在一起时所得的评价函数达到最大值(或最小值,依评价函数选取的不同,取最大或最小值),直到评价函数的值不再增加(或减小)时为止(亦或者达到指定的属性数为止)。这种算法的不足是:虽然考虑了所选属性与已入选属性之间的相关性,但却未考虑未入选属性之间的统计相关性;并且一旦某个属性已入选,即使由于后加入的属性使它变为多余,也无法再剔除。

广义逐步向前选择方法是逐步向前选择方法的推广。针对逐步向前选择方法"未能考虑未入选属性之间的统计相关性"的缺点[2],该方法每次从未入选的属性中挑选的不止是一个属性而是多个属性。广义逐步向前选择方法的缺点是计算量要比逐步向前选择方法大很多,并且也未解决"一旦某个属性已入选,即使由于后加入的属性使它变为多余,也无法再剔除"的问题。

2 无监督属性选择方法的思路

通过对传统的逐步选择方法的分析发现,它们的不足都是由属性之间的统计相关性引起的。那么,是否可以提出一种能够解决属性之间统计相关性的方法弥补上述的不足来进行属性的选择呢?在统计分析中,因子分析就具有这种性质。因此,笔者认为可以因子分析的思想为基础,来进行属性子集的选择。

假设一个事务性数据库中有 n 条记录,p 个属性,则

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

(1)建立属性的相关系数阵 $R=(r_{ij})_{p\times p}$

其中:$r_{ij}=\dfrac{\sum_{a=1}^{n}(x_{ai}-\bar{x}_i)(x_{aj}-\bar{x}_j)}{\sqrt{\sum_{a=1}^{n}(x_{ai}-\bar{x}_i)^2}\sqrt{\sum_{a=1}^{n}(x_{aj}-\bar{x}_j)^2}}=\dfrac{1}{n}\sum_{a=1}^{n}x_{ai}x_{aj}$

(2)求 R 的特征根及特征向量,分别记为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p>0$ 和 $u_1,u_2,\cdots,u_n$,则根据 $\sum_{i=1}^{m}\lambda_i/\sum_{i=1}^{p}\lambda_i\geqslant 85\%$,取前 m 个特征根及相应的特征向量形成矩阵 A,即

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}=\begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & & \vdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pm}\sqrt{\lambda_m} \end{bmatrix} \quad (2)$$

此时,矩阵 A 找出了能够控制所有属性的少数几个属性,并且得到了根据相关性大小进行的属性分类。这样的分类是同类内的属性之间相关性较高,不同类间的属性相关性

基金项目:国家统计局全国统计科学研究计划一般资助项目(2007LY010)

较低。所形成 m 个属性的分类 $\Psi_1,\Psi_2,\cdots,\Psi_m$，用来反映 m 个方面的内容，表示为 $F_1,F_2,\cdots,F_m$ 即为m 个因子。λ_i 在 $\sum_{i=1}^{m}\lambda_i$ 中的比重越大，能够说明 F_i 覆盖数据的自然分类的能力越强。

(3)每一个属性都可以表示为

$$X_i=a_{i1}F_1+a_{i2}F_2+\cdots+a_{ij}F_j+\cdots+a_{im}F_m+\varepsilon_i \quad (i=1,2,\cdots,p) \qquad (3)$$

可以得到，X_i 与 F_j 的协方差为：

$$Cov(X_i,F_j)=Cov(\sum_{k=1}^{m}a_{ik}F_k+\varepsilon_i,F_j)$$

$$=Cov(\sum_{k=1}^{m}a_{ik}F_k,F_j)+Cov(\varepsilon_i,F_j)=a_{ij} \qquad (4)$$

如果对 X_i 作了标准化处理，X_i 的标准差为 1，且 F_j 的标准差也为 1，因此

$$r_{X_i,F_j}=\frac{Cov(X_i,F_j)}{\sqrt{D(X_i)}\sqrt{D(F_j)}}=Cov(X_i,F_j)=a_{ij} \qquad (5)$$

对于标准化后的 X_i，a_{ij} 是 X_i 与 F_j 的相关系数，它一方面表示 X_i 对 F_j 的依赖程度，绝对值越大，密切程度越高；另一方面也反映了变量 X_i 对 F_j 的相对重要性。

由此，根据“无监督属性选择中属性子集最优的标准是该子集能够最好地覆盖数据的自然分类”的原则，属性子集选择的路径可以首先考虑 λ_i 在 $\sum_{i=1}^{m}\lambda_i$ 中的比重由大到小的顺序。其次，在 $\Psi_1,\Psi_2,\cdots,\Psi_m$ 的每个分类中，属性入选的顺序可以依 a_{ij} 的值，由高到低进行。

在此之前需要注意的是，由于在 $\Psi_1,\Psi_2,\cdots,\Psi_m$ 分类中，属性的相关性很高，为了避免将这些高度相关的属性都选入属性子集中，造成冗余，影响属性子集选择的效果，必需进行消除属性间相关性的处理。可以根据相关系数阵 $R=(r_{ij})_{p\times p}$，在 m 个属性的分类 $\Psi_1,\Psi_2,\cdots,\Psi_m$ 内分别将相关度高于某阀值的属性再次分组；然后，在这些组中仅留下最高的属性；再依 a_{ij} 的值，由高到低进行属性子集的选择。

3 逐步向前无监督属性子集选择方法的具体步骤

根据上述属性子集选择的思路得到的这种属性子集选择方法的具体步骤如下：

(1)计算矩阵 A，形成 m 个属性的分类 $\Psi_1,\Psi_2,\cdots,\Psi_m$。

(2)删除冗余。对每一类内的属性进行相关分析，将相关度大于等于阀值（一般大于等于 0.8）的属性分为一组记作 Ω_j，在 Ω_j 内留下因子载荷较高的属性，删掉该组内其它的属性。

(3)形成属性子集的核。按照 Ψ_1,Ψ_2 到 Ψ_m 的顺序依次选择同类内 a_{ij} 值最大的属性作为属性子集的核，称为 f_1。

(4)选择最佳子集。根据逐步向前选择算法，每次从剩下的属性中，选择使与已入选的属性组合在一起时所得的类内离差平方和达到最小值的属性加入到属性子集中；选择一直进行到类内离差平方和不再减少为止或者进行到指定的属性个数 d 为止。

4 统计模拟及方法验证

这里选用 2006 年各地区城镇居民家庭平均每人全年消费性支出样本集对上述方法进行模拟和验证。首先，计算 8 个属性的相关矩阵(见表 1)。

表 1　　8 个属性的相关阵

	食品	衣着	家庭设备用品及服务	医疗保健	交通和通信	教育文化娱乐服务	居住	杂项
食品	1.000	0.277	0.748	0.518	0.887	0.832	0.802	0.764
衣着	0.277	1.000	0.526	0.638	0.439	0.580	0.387	0.578
家庭设备用品及服务	0.748	0.526	1.000	0.718	0.767	0.943	0.769	0.833
医疗保健	0.518	0.638	0.718	1.000	0.578	0.733	0.699	0.704
交通和通信	0.887	0.439	0.767	0.578	1.000	0.870	0.762	0.777
教育文化娱乐服务	0.832	0.580	0.943	0.733	0.870	1.000	0.824	0.868
居住	0.802	0.387	0.769	0.699	0.762	0.824	1.000	0.811
杂项	0.764	0.578	0.833	0.704	0.777	0.868	0.811	1.000

资料来源：《中国统计年鉴 2007》。

然后，根据相关系数矩阵计算特征根及相应的特征向量。$\sum_{i=1}^{2}\lambda_i/\sum_{i=1}^{8}\lambda_i=86.376\%$，其中，$\lambda_i/\sum_{i=1}^{p}\lambda_i=74.815\%$。以上数据说明，$F_1$ 对 31 个省市自治区的分类有重大影响。计算矩阵 A 结果如表 2。

表 2　　矩阵 A 的计算结果

属性 / F	F_1	F_2
食品	0.954	0.097
衣着	0.142	0.933
家庭设备用品及服务	0.779	0.492
医疗保健	0.481	0.743
交通和通信	0.886	0.265
教育文化和娱乐服务	0.836	0.493
居住	0.847	0.320
杂项	0.768	0.511

这样，8 个属性被分为两类：“食品”、“交通和通信”、“居住”、“教育文化和娱乐服务”、“家庭设备用品及服务” 以及“杂项”为第一类，记作 Ψ_1，其中，“食品”的载荷最高；“衣着”和“医疗保健”为第二类记作 Ψ_2。本文将相关度的阀值定为 0.85。这样由表 1 及表 2 可知，Ψ_1 中产生的分组是：Ω_1 为“食品”和“交通和通信”；Ω_2 为“教育文化和娱乐服务”与“家庭设备用品及服务”、“杂项”；Ω_3 为居住。在 Ψ_2 中，“衣着”和“医疗保健”相关度仅为 0.638，因此各成一组。

接下来属性子集选择的步骤是：按照每一类内的高低及相关性，删除掉的属性分别是“家庭设备用品及服务”、“交通和通信”以及“杂项”；选取“食品”和“衣着”作为核，组成属性子集 f_1；剩下的属性中，依次添入子集 f_1 的属性是“居住”、“医疗保健”以及“教育文化娱乐服务”。其中，当属性子集由“衣着”、“食品”、“居住”、及“医疗保健”组成时，类内离差平方和就已达到最小，再继续添入属性时，类内离差平方和最小值不再改变。因此，最优的属性子集应由上述四个属性组成，这里用 f_2 表示。

基于流动性调整的期望损失研究

胡小平，吕宏生，何建敏
（东南大学 经济管理学院，南京 210096）

摘 要：与传统 VaR 方法相比，期望损失是一致风险测度，满足次可加性，但忽视了资产的流动性风险。流动性调整在险损失虽然考虑了流动性，但传统 VaR 方法的不足限制了 La-VaR 的实际应用范围和效果。鉴于此，基于相对半价差，文章研究了如何将流动性引入期望损失，得到带有流动性调整因素的期望损失模型，并给出了计算 La-ES 的仿真算法。实证分析表明，一致性风险测度 La-ES 既能够覆盖大部分极端风险，又表现的不太保守，是一种较好的风险度量工具。

关键词：期望损失；相对半价差；流动性调整；风险价值
中图分类号：F830.59 文献标识码：A 文章编号：1002-6487(2009)07-0038-04

0 引言

自 JP.摩根推出风险管理新产品 RiskMetrics 后，风险价值(VaR：Value at Risk)由于其概念简单、计算方便和易于使用等特点，已经成为金融风险管理的常用工具，巴塞尔委员会的《市场风险修正案》也把它作为市场风险的量度标准。德国有色金属公司在衍生品市场上的惨痛损失，和美国长期资本管理公司的破产都是由于资产的流动性出了问题。LTCMC 在破产前一直在使用传统的 VaR 风险模型作为公司的风险测量与控制工具，正是忽略了流动性风险，从而大大低估了资产所面临的整体风险，当极端风险事件发生时不可避免地走向破产。忽略流动性的教训使得监管机构和机构投资者认识到流动性风险在风险管理中的重要性，并制定新的风险管

基金项目：国家自然科学基金资助项目(70671025)

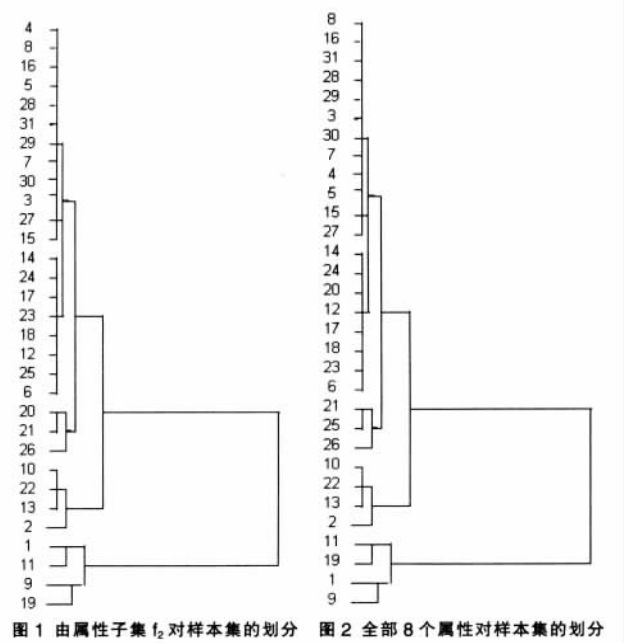

图 1 由属性子集 f_2 对样本集的划分 图 2 全部 8 个属性对样本集的划分

属性子集 f_2 对样本集的划分见图 1。图 1 将 31 个省市划分为三类：第一，“1”、“9”、“11”、“19”分别代表北京、上海、浙江和广东，这四个省市为一类；其余省市为一类。将图 1 的这种划分与全部 8 个属性对样本集划分的结果（见图 2）比较，可以发现，二者的分类结果完全一致，说明该属性子集具有对全部属性的概括能力，达到了“覆盖自然分类”的子集选择标准。

5 逐步向前无监督属性子集选择方法的优点和局限性

优点：该方法虽然未能解决“无法剔除已入选子集中的属性”的问题，但它既考虑到了所选属性与已入选属性间的相关性，又考虑了未入选属性之间的关系；并且极大程度地避免了“后入选的属性使子集中的属性变为多余”的情况发生；另外，由于在逐步向前选择算法之前先删除了一些属性，使得它的计算量相对减少了很多。

局限性：该方法对于 KMO 值小于 0.6 的样本集不适合；并且仍然是一种启发式的搜索方法，无法得到全局最优解。

参考文献：
[1]张莉，孙刚，郭军.基于 K-均值聚类的无监督特征选择方法[J].计算机应用研究，2005，(3).
[2]边肇祺，张学工.模式识别[M].北京：清华大学出版社，2005.

（责任编辑/易永生）