文章编号:1672 - 058X(2011)06 - 0602 - 06

R 软件的数据挖掘应用*

陈荣鑫
(集美大学 计算机工程学院,福建 厦门 361021)

摘 要:开源 R 软件集成了各种的数据分析和可视化方法,具备强大的数据分析功能和良好的可扩展性,适用于数据挖掘;结合城市主要经济指标的数据挖掘案例,给出了 R 软件在挖掘过程中各主要阶段的应用方法;数据准备阶段包括数据抽取、数据选择与统计分析应用;挖掘建模阶段给出了聚类和分类的典型挖掘应用;模型评估阶段给出了决策树的评估方法;从简洁的 R 语言脚本设计和良好的分析效果,展示了 R 软件的基本特点和在数据挖掘应用中的优势。

关键词:R 软件;数据准备;挖掘建模;模型评估
中图分类号:TP315 **文献标志码**: A

数据挖掘方法通过对数据的分析,发现有用的规律和概念,以提高数据拥有者对原始数据的深层次理解与认识,满足决策需求[1]。目前存在各种挖掘软件,主流的商用挖掘工具比如 Unica、SAS/EM、Insightful Miner、IBM IM 和 SPSS 等,这些软件特点是面向通用挖掘问题,功能较为完善,具备较好的性能。但一般都存在可扩展性不强、成本较高等缺点。开源软件能有效克服这些缺点,比较著名的包括 Weka、YALE、KNIME、Orange 和 R 等。R 软件是一款集成了数据操作、统计和可视化功能的优秀的开源软件[2]。R 软件具备高效的数据处理和存储功能,擅长数据矩阵操作,提供了大量适用于数据分析的工具,支持各种数据可视化输出。R 软件的一大优势是分析人员可利用简单的 R 程序语言描述处理过程,以构建强大的分析功能。此外,R 软件具备良好可扩展性,来自世界各地开源社区的研究者为其提供了各种丰富的工具包。由于 R 软件能结合各种挖掘算法,有效地简化数据分析过程,适用于数据挖掘领域。在此通过具体案例,探讨 R 软件在数据挖掘过程中各主要阶段的应用。

1 数据挖掘阶段

数据挖掘过程一般包括挖掘任务定义、数据准备、挖掘建模、模型评估和模型应用等阶段[3]。

(1) 任务定义。分析人员通过与挖掘系统交互,完成挖掘任务的定义。要求系统提供交互界面,并能生成任务描述信息。

(2) 数据准备。是挖掘的预处理阶段,包括数据抽取、数据集成、数据选择和数据转换等步骤。首先数据抽取把挖掘对象数据加载进入系统;数据整理用于删除噪声、不一致或重复的数据;数据选择用来抽取与分析任务相关的数据;数据转换则把数据转换或合并成适当形式,以利于挖掘的执行。

(3) 挖掘建模。根据已定义的挖掘任务,选择分类、聚类、关联规则等具体的挖掘方法进行建模。由于数据准备和数据挖掘本身都会涉及各种算法,然而每种算法有其解决特定问题的优势,又有其不适用于其他问题的劣势。显然,挖掘系统只有集成多种算法可供用户选择,才会有良好的实用性。

(4) 模型评估。对完成建模后的结果进行解释和评估,可采用可视化和用户易于理解的知识表示方式

收稿日期:2011 - 08 - 10;**修回日期**:2011 - 09 - 21.
* **基金项目**:福建省自然科学基金项目(2008J04005).
作者简介:陈荣鑫(1975 -),男,福建厦门人,讲师,硕士,从事软件自动化和数据库技术研究.

来表达挖掘结果。比如,采用图形化的决策树模型来表示分类模型,采用“if…then…”规则形式来表示关联模型。可视化效果对于提高挖掘结果的可解释性和知识的易理解性具有重要作用。

(5) 模型应用。发布通过评估的模型,提供用户模型应用服务。比如用户可应用已完成的分类模型对新实例进行类别预测。

2　数据准备

2.1　典型案例

采用的研究案例中,挖掘对象为我国 36 个省会城市和计划单列市的主要经济指标统计数据[4],据此拟对我国城市经济发展情况进行分析。数据如表 1 所示,原始数据来自中国统计年鉴,经过了简单处理,获得了城市发展各项指标的人均数据。各个属性说明如下:A_0 为城市名称,A_1 为年底人口总数(万人),A_2 为地区生产值(千元/人),A_3 为地方财政预算内收入(千元/人),A_4 为地方财政预算内支出(千元/人),A_5 为固定资产投资(千元/人),A_6 为城乡居民储蓄(千元/人),A_7 为 社会商品零售(千元/人),A_8 为货物进出口(千美元/人),A_9 为普通高等学校在校学生数(人/百人),A_{10}为医院卫生院(所/万人),A_{11}为执业医师(人/百人)。统计数据共有 36 条记录,由于篇幅所限,表 1 中仅列出其中 4 条实例信息记录。

表 1　城市经济指标人均数据

A_0	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}
北京	1 246	9.8	16.3	18.6	39	116.9	42.6	17.2	5	0.5	50
天津	980	7.7	8.4	8.4	51.1	50.7	24.8	6.5	4	0.5	28
石家庄	977	3.1	1.3	1.3	24.9	26.3	12.2	0.6	4	0.4	21
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
乌鲁木齐	241	4.5	4.7	4.7	17.1	43.2	19.6	1.5	5	0.7	422.2

2.2　数据抽取

作为挖掘对象的数据存储形式多样,一般有文件、数据流和数据库等形式存储,挖掘分析则在计算机内存中进行,因此第一步需进行数据抽取,获取外部数据到内存中。R 软件提供了多种数据文件和数据库的存取方法。

数据文件包括通用型文件如纯文本文件、Excel 等。例如读取 csv 格式数据的方法用 dataset < - read. csv("CityData. csv",row. names = 1),其中 row. names 约定了数据表的第一行为表头,包含了各属性名称;而存储 csv 数据的方法是 write. csv( )。通过加载 foreign 功能模块,R 软件还支持存取 SAS、SPSS、S - PLUS 等专业软件所使用的数据文件。比如 dataset < - read. spss("test. sav", to. data. frame = TRUE)用于读取 SPSS 数据文件,第二个参数表示把读入的数据形成数据框(data frame)。

R 软件同样提供了各种主流数据库的连接方法。比如通过安装 RODBC,获取对 ODBC 数据源的访问能力;RJDBC 用于访问 JDBC 数据源;ROracle 用于访问 Oracle 数据库;RMySQL 用于访问开源数据库 MySQL。由于实际应用中往往需要对异构数据源进行挖掘,R 软件提供了针对各种数据源的访问接口,而系统内部有统一的数据表达形式,因此适用性很强。

挖掘对象数据多数为关系类型,可以表示为矩阵形式。R 软件中对矩阵的存储采用数据框,其特点是每类代表一个属性变量,各列可以是不同类型数据,而每行是一个对象实例。数据框是一种扩展了的矩阵,可采用对矩阵的下标引用方法来引用其元素或子集,方便各种数据的存取操作。

2.3　数据选择

为了寻找合理的挖掘对象数据集,缩小处理范围,提高数据挖掘的质量和效率,需要进行数据选择。数

据选择使得后续的挖掘工作聚焦到与挖掘任务相关的数据子集中。一般可通过属性选择和数据采样进行选择过滤。

属性选择能保证数据挖掘的实效性,通常需要结合专业背景知识完成。表1中,除了 A_0 属性作为实例标记,由于研究的内容是人均数据,A_1 属性显然也需要排除,剩下部分可作为选择属性。R 程序对属性选择描述非常直接,只要在相应的向量构造函数 c()中选取属性即可,可以是属性名称也可以是属性顺序号。比如 subset < - subset(dataset, select = c(2:11))语句,用属性顺序号描述了从 $A_2 \sim A_{11}$ 共10个连续属性的选择。此外,可考虑通过主成分分析等手段,对选择属性进一步简化,这要结合考虑后续选用的挖掘算法特点。

挖掘过程中,通常需要用数据采样获得部分实例以进行模型训练,再用剩余的部分实例进行模型验证。对于大型数据集的挖掘,同样可通过数据采样获取数据子集以完成挖掘建模,满足了内存或算法的限制,在保持足够精度的前提下提高了挖掘效率。在 R 程序语句 sm < - sample(nrow(dataset), 25)中,nrow(dataset)获得数据集 dataset 的实例数 N,用 sample 函数随机获取 $1 \sim N$ 区间中的25个整数作为样本 id。再用 data = dataset[sm, c(2:12)]语句,从数据框 dataset 中根据 sm 中的样本 id 进行数据选择。

2.4 数据统计分析

通过统计分析挖掘对象数据,深入理解数据,对完善和细化挖掘任务有很大帮助。统计分析能获得数据基本特征,包括数据的集中位置、分散程度和分布情况等,而统计结果可进行直观展示,比如采用直方图、经验分布图、QQ 图和茎叶图等方式进行数据分布的可视化输出[5]。

R 软件提供了大量统计函数和绘图函数支持统计分析。以下程序分析了地区平均生产值的分布情况并进行可视化,结果如图1,其中横坐标为各人均产值区间,纵坐标为概率值。从图1中发现多数城市的人均产值在3~4千元/人区间,总体密度估计并未呈正态分布。

```
pd < - CityDataMYM 地区生产值   #获取数据框的一个列
hist(pd, freq = FALSE)   #绘制直方图
lines(density(pd), col = "red")   #绘制密度估计曲线
```

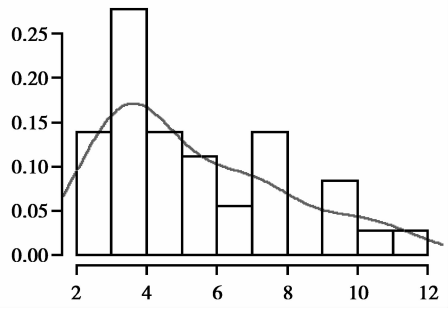

图1 直方图分析

3 挖掘建模

3.1 聚类挖掘

分析城市经济发展的类型时,在初始数据处理过程中,由于数据实例的内部聚合关系尚不明确,没有一个已知的分类准则,故采用聚类这种非监督的建模方法,尝试对城市发展类型进行区分,以获取类别信息。聚类分析对相似实例进行分组,而相似性标准则是采用计算距离的方法。R 软件具备数种聚类挖掘算法,包括典型的基于层次的聚类如系统聚类法,以及基于划分的聚类如 k-means 等方法。

(1) 基于层次的聚类。系统聚类法是一种基于层次的聚类,使用广泛,其基本思想是:初始将各个样本各自作为一类,规定样本之间距离和类之间的距离,然后将距离最近的两个类合并为一个新类,继续计算新类与其他类的聚类,重复进行最近两个类的合并工作,直至所有样本合并为一类,根据类合并顺序,自下而上,自然形成了层次关系。

R 软件提供了 hclust()函数用于层次聚类。hclust()函数形如 hclust(d, method = "ward", members = NULL),其中 d 是距离结构数据,method 用于指定聚类的方法,可选用最短距离法、最长距离法、中间距离法、Mcquitty 相似法、类平均法、重心法或离差平方和法。考虑到样本数量和对聚类精度的要求,选用离差平方和法。离差平方和法即著名的 Ward 方法,是一种基于方差分析的方法,基本思想是如果聚类区分得正确,那么同类样本之间的离差平方和应当较小,而不同类样本之间的离差平方和应当较大。

由于聚类是非监督的,类个数的合理性值得关注。根据实际应用情况进行调整,最后指定产生4个聚类,对城市经济指标进行系统聚类并绘图的程序如下。

```
subset < - subset(dataset, select = c(2:11))
```

```
city < - dist(subset)　#计算 Euclide 距离
hclust < - hclust(city, "ward")　#进行 Ward 法系统聚类
plclust(hclust)　#绘制系统聚类的谱系图
result < - rect.hclust(hclust, k = 4, border = "red")　#指定 4 个聚类的划分.
```

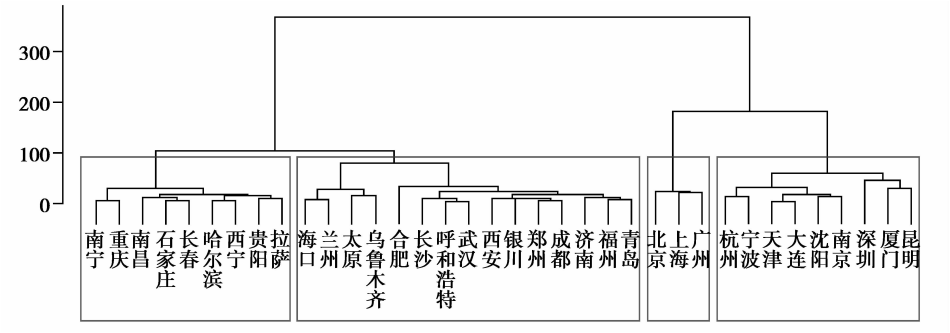

图 2　系统聚类结果

根据图 2 所示的聚类结果,可获得 4 个聚类的具体划分。比如北京、上海和广州属于同一发展类型。

(2) 基于划分的聚类。*k*-means 是经典的基于划分的聚类方法,其基本思想是使聚类性能指标最小化。所用的聚类准则函数是聚类集中每个样本点到该类中心的距离平方之和,应使其最小化。为此,首先根据给定聚类数 K,为每个聚类确定一个初始聚类中心;其次将样本集里的各个样本按最小距离原则分配到最邻近的聚类,并使用每个聚类中的样本均值作为新的聚类中心,如此重复直到聚类中心不发生变化;最后可获得 K 个聚类。

R 软件中 *k*-means 方法的函数形式如:*k*means(*x*, centers, iter. max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen")),其中 *x* 是挖掘对象的数据矩阵;centers 用于指定聚类个数或初始聚类中心;iter. max 为最大迭代数,默认为 10;如果 centers 指定了聚类个数,则 nstart 表示选用的随机集个数;algorithm 表示采用的具体算法。聚类可视化采用判别投影绘制函数 plotcluster,把数据对象映射到平面空间,展示聚类之间的异构性。对城市经济指标进行 *k*-means 聚类的程序如下:

```
kmeans < -kmeans(dataset[ ,c(2:11)], 4)
library(fpc) #引入 fpc 包以支持聚类图的绘制
plotcluster(dataset[ ,c(2:11)], kmeansMYMcluster)
```

根据图 3 所示的聚类结果,可获得 4 个聚类的具体划分。从 kmeansMYMcluster 数据查看各个实例所属的聚类,发现该法获得与系统聚类类似的结果,比如北京、上海和广州也属于同一类发展类型,对应图 3 中右上角落的 3 个点。

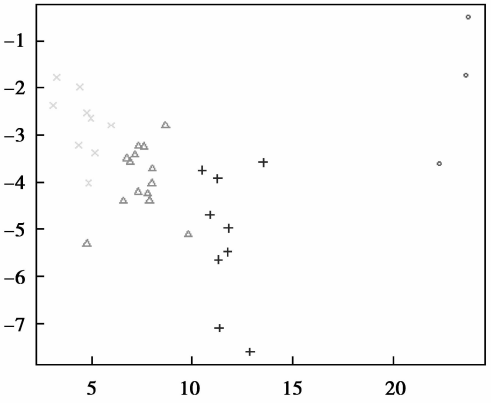

图 3　*k*-means 聚类结果

3.2　分类挖掘

分类挖掘根据选定的目标属性,对数据实例挖掘建模以获取决策树模型,该模型用于分类预测。通过决策树模型,容易进一步获取类别决策规则。分类回归树是一种重要的决策树方法,其基本思想是通过二分递归划分数据,直到满足特定终止条件停止树的生长。终止条件可以是节点已获得分类值,或者树深度达到用户指定值,或者节点中样本数少于用户指定值等;选择的划分点能使各划分部分均值的方差和最小。

R 软件提供的分类回归树函数形如:rpart(formula, data, weights, subset, na. action = na. rpart, method, model = FALSE, *x* = FALSE, *y* = TRUE, parms, control, cost, ……),其中 formula 是回归方程,如 $A \sim A_1 + A_2 + A_3$,左式 A 为目标属性,右式为属性列表;data 为待分析数据;control 用于控制决策树生成的细节。根据 3.1 节聚类分析的结果,获得城市经济发展类型,据此作为原有数据集新增的一个属性,该属性作为分类决策目标;值域是 1 ~4 的离散值,表示四种发展类型。对数据进行采样建模,分类决策树的构造程序如下。

```
library(rpart)　#引入 rpart 包以支持分类
```

```
sample < - sample(nrow(dataset), 25)  #采样率 70% 时,用 25 个实例建模
rpart < - rpart(发展类型 ~ ., data = dataset[sample,c(2:12)], method = "class", control = rpart. control (minsplit = 1))
# minsplit 规定了每节点上最少出现的实例次数
plot (rpart, uniform = T, branch = 0.4 ,compress = T)  #决策树的可视化输出
text (rpart , use. n = T)
```

图 4 为获得的决策树,每个叶节点标示了分类目标属性值,带斜杠的数据表示属于各目标属性值的实例个数。为了简化获得的决策树模型,避免产生过度拟合,需要进一步分析进行剪枝[6]。考察 rpartMYMcptable 变量里各分裂点的误差、标准差信息,在保证预测误差尽量小的前提下,通过控制树的复杂性,使得树的规模较为合理。设置复杂性参数 cp 为 0.06,剪枝程序如下,获得剪枝后的决策树如图 4。

```
rpartp < - prune(rpart, cp = 0.06)
plot (rpartp, uniform = T, branch = 0.4 ,compress = T)
text (rpartp , use. n = T)
```

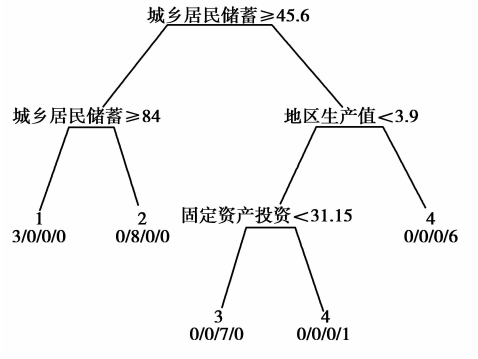

图 4 分类决策树

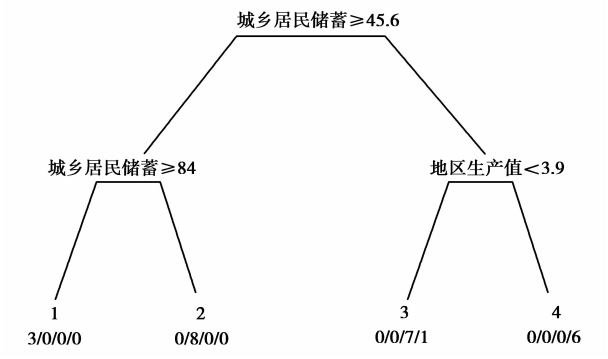

图 5 剪枝后的决策树

通过剪枝后决策树已经足够简化,由此产生的决策规则将十分简单。图 5 对应的规则如下:

Rule 1: if (城乡居民储蓄 < 45.6 and 地区生产值 > =3.9) then 4;

Rule 2: if (城乡居民储蓄 < 45.6 and 地区生产值 < 3.9) then 3;

Rule 3: if (城乡居民储蓄 > =45.6 and 城乡居民储蓄 < 84) then 2;

Rule 4: if (城乡居民储蓄 > =84) then 1。

4 模型评估

对分类决策树模型进行验证,使用采样剩余的 30% 数据,即总共 11 个数据记录作为检验模型的预测实例。编写 R 程序如下:

```
predict < - predict(rpartp, dataset[ - sample, c(2:12)], type = "class")
table(predict, dataset[ - sample, c(2:12)]MYM 发展类型, dnn = c("预测值", "实际值"))
round(100 * table(predict, dataset[ - sample, c(2:12)]MYM 发展类型, dnn = c("预测值", "实际值"))/length(predict))
```

获得的评估矩阵如表 2 所示,左边为预测结果实例分布矩阵,右边为对应百分比矩阵,可见该模型在本测试中正确率达 91% 左右。

表 2 模型评估结果矩阵

实例分布						百分比					
		实际值						实际值			
		1	2	3	4			1	2	3	4
预测值	1	0	0	0	0	预测值	1	0	0	0	0
	2	0	3	0	0		2	0	27	0	0
	3	0	0	4	1		3	0	0	36	9
	4	0	0	0	3		4	0	0	0	27

5　结束语

一个实用的数据挖掘系统,一方面需要具备完善的挖掘功能,另一方面需要有友好的用户界面,R 软件具备构建实用性强的数据挖掘系统的各种条件。用户可通过简洁的 R 程序语言开发各种算法,增强 R 软件的挖掘能力;由于 R 软件的开源特点,容易获得大量优秀的功能包,完善挖掘功能;R 软件提供了其他高级语言如 C 语言的编程接口,容易与其它挖掘软件实现互操作。R 软件的主要交互手段是命令行界面,通过编写 R 程序脚本来调用分析功能,目前出现了几种图形交互界面,比如 Rattle[7] 和 R Commander[8] 等,极大方便了挖掘的实施。图形界面除了需要改善数据可视化效果,还需要支持流程化的挖掘任务定义和执行。通过功能的增强和用户界面的改进,相信 R 软件在未来的数据挖掘领域中将得到更为广泛的应用。

参考文献:

[1] HAN J, KAMBER M. 数据挖掘:概念与技术[M]. 北京:机械工业出版社, 2001
[2] VENABLES W N, SMITH D M. R Development Core Team. An introduction to R [EB/OL]. (2011-04-13) [2011-05-01]. http://cran. r-project. org/doc/manuals.
[3] 邵峰晶, 于忠清. 数据挖掘原理与算法[M]. 北京:中国水利水电出版社, 2003
[4] 中国资讯行. 中国 2009 年省会城市和计划单列市主要经济指标统计(中国统计年鉴)[EB/OL]. (2009-12-31) [2011-05-01]. http://www. bjinfobank. com
[5] 薛毅, 陈立萍. 统计建模与 R 软件[M]. 北京:清华大学出版社, 2007
[6] 谢益辉. 基于 R 软件 rpart 包的分类与回归树应用[J]. 统计与信息论坛, 2007, 22(5): 67-70
[7] WILLIAMS G. Rattle: a data mining GUI for R [J]. The R Journal, 2009(1): 45-55
[8] FOX J. Getting Started With the R Commander: A Basic-Statistics Graphical User Interface to R [J]. Journal of Statistical Software, 2005, 14(9): 1-42

Data Mining Application Based on R

CHEN Rong-xin

(Computer Engineering College, Jimei University, Fujian Xiamen 361021, China)

Abstract: R is open source software integrated with various data analysis and visualization methods. It has powerful data analysis ability and good extendibility; therefore it is adapted to data mining. Through the cities' major economic indicators of mining case, the application methods are presented to complete the main data mining procedures. Data preparation includes data extraction, selection and statistic analysis; mining modeling includes cluster and classification application; model evaluation includes the assessing approach for decision tree. From the concise R script design style and excellent analysis effects, the general features of R and its application advantage in data mining are revealed.

Key words: R; data preparation; mining modeling; model evaluation

责任编辑:代小红