综述与评论 Review and Comment

基于 SPSS 和 KNIME 的 K-means 聚类结果研究

陈 朋
(上海大学 管理学院,上海 200444)

摘 要: 分别采用 SPSS 和 KNIME 软件分析了大样本和小样本两种数据集来比较不同的分析工具在运用 K-means 算法后得到的结论以及它们之间存在的差别,以期望对数据挖掘工具的选择带来指导作用。

关键词: 数据挖掘; K-平均; 聚类分析; SPSS
中图分类号: TP311 **文献标识码**: A **文章编码**: 1674-7720(2010)12-0001-03

Analysis of the K-means cluster results based on SPSS and KNIME

CHEN Peng
(School of Management, Shanghai University, Shanghai 200444, China)

Abstract: In this paper, two kinds of data sets are analyzed by using SPSS and KNIME software. The purpose is to compare the outcomes after using K-means algorithm, as well as the differences between them. It is expected that the conclusions of the paper will play a guiding role in the choice of the data mining tools.

Key words: data mining; K-means; cluster analysis; SPSS

近年来,随着数据挖掘工具的不断涌现,如何选择数据挖掘工具已成为数据挖掘技术引入的一大难题。Elder Research Inc.(ERI) 提供了许多数据挖掘系统和工具的性能测试报告[1]。

不同的软件有不同的特点,一方面要关注它们的性能,同时也要关注这些软件应用到数据挖掘中时产生的结果。下面仅讨论在 SPSS(ver.16)和 KNIME(ver.2.1.1)中 K-means 算法所产生的结果的特点和差别。

1 K-means 算法

K-means[2]算法是一种著名的并且常用的聚类方法。K-means 以 k 为参数,把 n 个对象分为 k 个簇(cluster),以使簇内具有较高的相似度,而簇间的相似度较低。相似度的计算是根据一个簇中对象的平均值(被看作簇的重心)来进行的。

K-means 算法随机地选择 k 个对象,每个对象初始地代表了一个簇的平均值或中心,根据其与各个簇中心的距离,对剩余的每个对象赋给最近的簇,然后重新计算每个簇的平均值。这个过程不断重复,直到准则函数收敛。通常,采用平方误差准则定义如下:

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|^2 \quad (1)$$

式中,E 是数据库中所有对象的平方误差的总和,p 是空间中的点,表示给定的数据对象,m_i 是簇 C_i 的平均值(p 和 m_i 都是多维的)。这个准则试图使生成的结果簇尽可能地紧凑和独立。

该算法尝试找出使平方误差函数值最小的 k 个划分。若结果簇密集,而簇与簇之间的区别明显时,其效果好。对处理大数据集,该算法的可伸缩度和效率相对较高,因为它的复杂度是 $O(nkt)$,其中,n 是所有对象的数目,k 是簇的数目,t 是迭代的次数。通常,$k<<n$,且 $t<<n$。这个算法经常以局部最优结束。

但是,K-means 方法只有在簇的平均值被定义的情况下才能使用,这可能不适用于某些应用,例如涉及有分类属性的数据。该方法的一个缺点是要求用户必须事先给出 k(要生成的簇的数目)。K-means 方法不适用于发现非凸面形状的簇。此外,它对于"噪声"和孤立点数据敏感,少量的该种数据能够对平均值产生极大的影响。

2 SPSS 和 KNIME 的介绍

SPSS(Statistical Product and Service Solutions)是世界上最早采用图形菜单驱动界面的统计软件,它最突出的特点就是操作界面极为友好,输出结果美观漂亮。它将几乎所有的功能都以统一、规范的界面展现出来,使用 Windows 的窗口方式展示各种管理和分析数据方法的功能,使用对话框展示出各种功能选择项,用户只要掌握

一定的 Windows 操作技能，粗通统计分析原理，就可以使用该软件为特定的科研工作服务。SPSS 采用类似 Excel 表格的方式输入与管理数据，数据接口较为通用，能方便地从其他数据库中读取数据。其统计过程包括了常用的、较为成熟的统计过程，完全可以满足非统计专业人士的工作需要。输出结果十分美观，存储时则是专用的 SPO 格式，可以转存为 HTML 格式和文本格式。对于熟悉老版本编程运行方式的用户，SPSS 还特别设计了语法生成窗口，用户只需在菜单中选好各个选项，然后按“粘贴”按钮就可以自动生成标准的 SPSS 程序，极大地方便了中、高级用户。

KNIME(Konstanz Information Miner)[3]数据挖掘工具基于 Eclipse 开发环境精心开发，可以扩展使用 Weka 中的挖掘算法。它采用类似数据流(data flow)的方式来建立分析挖掘流程。挖掘流程由一系列功能节点(node)组成，每个节点有输入/输出端口(port)，用于接收数据或模型、导出结果。节点之间的连接很方便，直接用鼠标拖拽连接端口即可。

KNIME 中每个节点都带有交通信号灯，用于指示该节点的状态(未连接、未配置、缺乏输入数据时为红灯，准备执行为黄灯，执行完毕后为绿灯)。KNIME 的特色功能 HiLite 允许用户在节点结果中标记感兴趣的记录，并进一步展开后续探索。

3 两类数据集的聚类结果

3.1 大样本数据集的聚类分析

鸢尾花数据集[4]被公认为最著名的用于数据挖掘的数据集，它包含 3 种植物种类，每种各有 50 个样本，所以样本总数为 150。它有花萼长(sepallength)、花萼宽(sepalwidth)、花瓣长(petallength)、花瓣宽(petalwidth)4 个属性，且都是数值属性。为了使实验结果更明显，添加属性 class 用以显示鸢尾花所处的类。数据库内的三个品种分别是 Iris-versicolor、Iris-setosa 和 Iris-virginicia。

设置 $k=3$，在 SPSS 中 K-means 的运算结果如表 1 所示，KNIME 中 K-means 的运算结果如图 1 所示。

表 1 SPSS 中 K-means 运算结果($k=3$)

Cluster	1	2	3
花萼长	6.9	5.0	5.9
花萼宽	3.1	3.4	2.7
花瓣长	5.7	1.5	4.4
花瓣宽	2.1	0.2	1.4
包含数目	38	50	62

由于两个工具得到的聚类序号不具有一一对应关系，为了比较两种结果需要对类号进行统一。根据鸢尾花数据集的属性 class 可以完成这项工作，得到统一后的结果如表 2 所示。

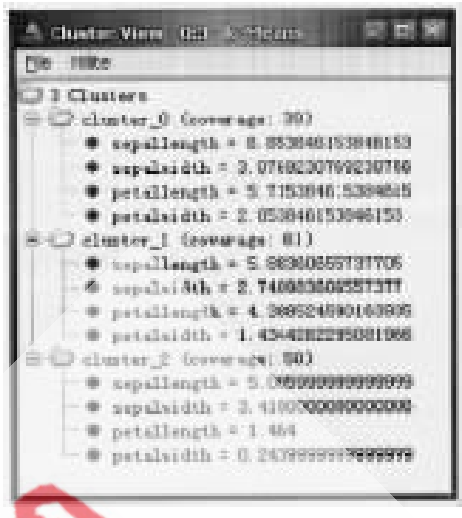

图 1 KNIME 中 K-means 运算的结果($k=3$)

表 2 进行统一后的聚类结果

CLUSTER	SPSS	KNIME
1	62	61
2	50	50
3	38	39

3.2 小样本数据集的聚类分析

实验所采用的小样本数据集如表 3 所示，当 k 分别设置为 2 和 3 时，使用 KNIME 的运算结果如图 2 所示，使用 SPSS 的运算结果如表 4 和表 5 所示。

表 3 数据准备

公 司	组织文化	组织氛围	领导角色	员工发展
Microsof	80	85	75	90
IBM	85	85	90	90
Dell	85	85	85	60
Apple	90	90	75	90
联想	99	98	78	80
NPP	88	89	89	90
北京电子	79	80	95	97
清华紫光	89	78	81	82
北大方正	75	78	95	96
TCL	60	65	85	88
娃哈哈	79	87	50	51
Angel	75	76	88	89
Hussar	60	56	89	90
世纪飞扬	100	100	85	84
Vinda	61	64	89	60

通过以上结果可以得出这样的结论：当 K-means 聚类方法用于大样本数据集的时候，无论是从聚类的最终中心距离，还是每个聚类内所有样本的数目，SPSS 和 KNIME 产生的聚类几乎是一致的，即排除 K-means 算法本身的一些缺陷以外，不同的分析工具不会对聚类结果产生显著的影响。

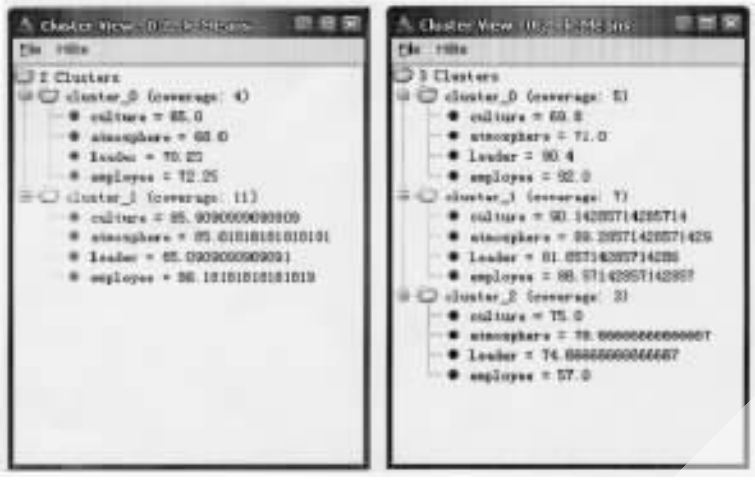

图 2 KNIME 中 K-means 运算的结果(k 分别取 2、3)

表 4 SPSS 中 K-means 运算的结果(k=2)

Cluster	1	2
组织文化	75.20	90.60
组织氛围	75.60	92.00
领导角色	87.60	74.60
员工发展	87.20	73.00
包含数目	10	5

表 5 SPSS 中 K-means 运算的结果(k=3)

Cluster	1	2	3
组织文化	89.50	68.33	79.00
组织氛围	88.75	69.83	87.00
领导角色	82.25	90.17	50.00
员工发展	83.25	86.67	51.00
包含数目	8	6	1

当 K-means 聚类方法用于大样本数据集的时候，本文只是选取了在 k 值分别是 2 和 3 两种情况下的结果，可以看出 k 值从 2 变为 3 时，两种工具得出的最终结果存在较大的差异，这是因为算法开始时随机地选择 k 个对象，两种工具选取的样本将会有所不同，加之样本总数的限制，最终导致了聚类结果的差别。

数据挖掘的工具还有很多，如何进行数据集的选取是目前存在的问题，为了得到更加完善和具有普通意义的结论，需要关注这些问题。

参考文献

[1] ABBOTT D W, MATKOVSKY I P. An evaluation of high-end data mining tools for fraud detection [R]. IEEE International Conference on Systems. 1998,12:12-14.

[2] HAN Jia Wei, KAMBER M 著.数据挖掘概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社, 2004.

[3] BERTHOLD M R, CEBRON N. Knime: the konstanz information miner[R]. http://www.knime.org, 2007,9.

[4] FISHER R A. Iris plants database[DB]. http://archive.ics.uci.edu/ml/datasets/Iris, 1988,7.

[5] 黄颖,李伟. EM 算法与 K-Means 算法比较[J]. 计算机与现代化, 2007(9):12-14.

[6] LI Tao Ying, CHEN Yan. An effective algorithm to solve optimal k value of K-means algorithm[S]. International symposium on distributed computing and applications to business, 2006.

[7] JIANG Sheng Yi, LI Qing Hua. An enhanced K-means clustering algorithm[J]. Computer Engineer & Science, 2006.

[8] 袁方,周志永,宋鑫.初始聚类中心优化的 K-means 算法[J].计算机工程, 2007,2(33).

[9] 张雪英. 国外先进数据挖掘工具的比较分析[J].计算机工程, 2003,9(29).

(收稿日期:2010-02-10)

作者简介:

陈朋,男,1987 年生,硕士在读,主要研究方向:数据挖掘。