基于校园一卡通消费数据的几种聚类算法的分析比较①

董新科, 张 晖

(西南科技大学 计算机科学与技术学院, 绵阳 621000)

摘 要: 随着高校管理信息化的加速和高校管理部门对各类校园信息资源需求的不断加强, 校园一卡通被广泛应用于学生生活的各个领域, 并要求对其存储的海量数据进行挖掘分析为各个部门提供决策依据. 聚类算法作为最常用的数据挖掘方法之一被广泛应用于一卡通数据挖掘, 但目前不清楚哪种方法更适用于一卡通数据. 使用多种常用聚类算法对一卡通数据进行了实验, 得出了最适合挖掘该数据的聚类算法, 并分析了相关原因.

关键词: 数据挖掘; 聚类; 高校消费数据; 校园一卡通

Analysis and Comparison of Several Clustering Algorithms Based on Campus Card Consumption Data

DONG Xin-Ke, ZHANG Hui

(School of Science and Technology, Mianyang 621000, China)

Abstract: With the acceleration of information technology in university management and the demand of university information continues to strengthen, the campus card is widely used in all aspects of student life, and requires data mining analysis for each sector basis for decision making. As one of the most popular data mining technologies, clustering algorithms are widely used for campus card consumption data mining. Howerve, people don't know which clustering algorithm is the most suitable for campus card data. This paper carries out experiments with multiple widely used clustering algorithms, obtaining the most appropriate data mining clustering algorithm for the data, and analyzes the reason.

Key words: data mining; clustering; university consumption data; campus card

1 引言

近年来, 校园信息化建设飞速发展, 一卡通等校园信息化软件得到了快速的实施. 这些信息化软件不但极大地方便了人们的生活, 而且积累了海量的信息. 对这些数据深入进行挖掘可以为教学、科研、后勤和管理等多个领域做出十分有益的贡献.

本文在前人研究的基础上, 从校园一卡通的消费数据入手, 通过常用的聚类算法对其进行分析, 采用多个衡量指标衡量这些算法在此数据上的聚类效果. 通过实验对比, 得出了最合适做校园一卡通消费数据挖掘的聚类算法, 为开发针对一卡通数据的数据挖掘系统奠定了基础.

2 相关工作

2009 年, 张兵兵[1]提出了数据挖掘技术在校园一卡通系统中的应用流程、方法和作用. 2010 年, 罗华群等[2]研究了校园卡数据库的数据筛选、聚类、关联方法, 并在此基础上分析和验证了学生生活和学习的关系. 2010 年, 李珊娜[3]以北京交通大学校园一卡通平台上设计和研发的学生行为分析系统和就餐消费分析系统为例, 介绍了对校园一卡通系统的数据挖掘的初步探索结果. 2012 年, 徐剑[4]等通过 K-means 算法对学生对热水使用情况进行了分析, 了解学生对热水的需求量, 从而提供给学校后勤部门一些参考意见.

从这些相关研究可以看出方法对我们目前的研究具

① 收稿时间:2013-07-16;收到修改稿时间:2013-09-22

有非常重要的作用, 聚类算法是一卡通数据挖掘的基础方法. 但是使用什么样的聚类算法对校园一卡通数据进行数据挖掘更合适这个问题并不清楚, 也没有进行相关说明. 本文将从一卡通数据挖掘的实际需要出发, 分析得出最适合一卡通消费数据挖掘的聚类算法.

3 常用聚类算法

Weka 是 Waikato 大学研究的开放源码的数据挖掘平台, 其中集成了大量能承担数据挖掘任务的机器学算法, 包括对数据进行预处理、关联规则挖掘、分类、聚类等[5]. 以 Weka 作为的实验平台, 并以 Weka 附带的聚类算法作为比较、分析、衡量的对象.

3.1 Weka 常用聚类算法简介

(1) Cobweb: Cobweb 算法是一个流行的简单增量聚类算法[6]. 它不仅能够聚类, 而且能更进一步找出每一个类的特征描述. 它假设在每个属性上的概率分布是彼此独立的.

(2) DBScan: 首先扫描数据库, 记录每一个点(记录)的 e—邻居个数, 如果一个记录的 e—邻居个数大于一个阈限值, 就这个记录叫做中心记录. 这样一个新的以这个记录为中心的类就产生了. 该方法可以用来过滤噪声数据, 可以快速发现任意形状的类[6].

(3) EM: EM[7]算法可被看作为一个逐次逼近算法, 包含两个步骤: E 步骤—计算期望值, M 步骤—重新计算参数值. 主要目的是提供一个简单的迭代算法计算后验密度函数.

(4) FarthestFirst: FarthestFirst 算法是快速的近似 k 均值的聚类算法, 是一种分层聚类算法. 有广度优先遍历、深度优先遍历两种形式.

(5) FilteredClusterer: FilteredClusterer 算法是一个综合的方法, 可以采用任何的过滤器进行数据的过滤, 同时可以采用任意的方法进行聚类.

(6) HierarchicalClusterer: HierarchicalClusterer 算法是一个层次聚类算法. 它产生一个嵌套聚类的层次, 算法最多包含 N 步, 在第 t 步, 执行的操作就是在前 t-1 步的聚类基础上生成新聚类.

(7) MakeDensityBasedClusterer: MakeDensityBased Clusterer[8]算法首先初始化一个没有子种群的全局种群, 再在全局种群中采用迭代搜索, 并对其中的个体进行聚类. 当聚类簇中的个体数目达到规定的最小规模时形成一个子种群, 然后在各子种群中进行迭代搜索并重新进行聚类,从而提高进化过程中种群的多样性, 增强算法跳出局部最优的能力.

(8) OPTICS: OPTICS 算法克服参数设置由用户决定的缺点, 并不显式地产生数据集聚类, 而是生成自动和交互的聚类结构. 它包含的信息等价于从一个广泛的参数设置所获得的基于密度的聚类. 簇排序可以用来提取基本的聚类信息(如簇中心, 任意形状簇), 也可以提供内在的聚类结构. 每个对象存储两个值: 核心距离(core-distance)和可达距离(reachability- distance).

(9) sIB 算法将待分析的数据对象按照其与另一数据对象的相关性进行“硬”划分, 使得划分在一起的对象充分体现出源数据对象蕴含的某个特征模式.

(10) SimpleKMeans 算法接受输入量 k, 然后将 n 个数据对象划分为 k 个聚类以便使得所获得的聚类满足: 同一聚类中的对象相似度较高; 而不同聚类中的对象相似度较小.

(11) Xmeans: XMeans 算法是 Kmeans 的改进, 在总体记录中通过 Kmeans 产生聚类, 再分别对每个聚类进行 Kmeans 式的迭代, 将某些子类再进行聚类, 直到达到用户设定的迭代次数为止.

3.2 Weka 常用聚类算法理论比较

深入研究上述聚类算法的特点和原理, 将 Weka 聚类算法进行对比得出如下的结论, 对比结果如表 1.

总的来说上述聚类算法各有优劣, 哪种算法更加符合本文的目标需求需要进行实验进行验证.

4 基于校园一卡通消费数据实验

本文实验是在校园一卡通消费数据的基础上, 使用 weka 软件进行分析, 采用了 weka 自带的常用的经典聚类算法. 同时, 根据一卡通数据挖掘的实际需求, 从多个维度分析了这些聚类算法在挖掘一卡通消费数据中的有效性和可用性.

4.1 实验数据选择和预处理

在西南科技大学数据中心中存储着大量的学生日常一卡通各种行为数据, 首先需要从中心库导出相关的数据进行清洗和集成. 通过人工筛选和处理, 获得了真实的学生日常消费数据.

针对大量的消费数据, 选择一学期内以下四个指标作为分析的依据: 消费总次数(xfzcs)、消费总金额(xfzje)、次均消费金额(cjxfje)和人均消费金额(rjxfje).

4.2 实验参数控制

为了使所选择的算法参数更加符合实际情况的需要,

需要对相关的参数进行设置, 以便达到最优的效果.

根据前期的问卷调查, 得出: 按照学生在学校的消费行为可以将该群体大致分为四大类. 因此, 聚类的数目 numClusters 参数选择 4 比较合适.

为了保证实验的效果, 聚类的最大迭代次数尽量选择大一些, 本文最大迭代次数参数 maxIterations 统一设为 500. 对于所有的实验数据采用将数据集 50%的数据用于训练, 50%的数据用于测试的方式进行实验.

表 1　聚类算法对比表

聚类算法	参数个数	其他聚类	优点	缺点	其他特点
Cobweb	3	N	1.自动修正划分中类的数目	1.基于属性独立的假设，易造成时间和空间的复杂性 2.不适于聚类大型数据库数据	对记录的顺序很敏感
DBScan	2	N	1.形成的簇可以有任意的形状 2.效率高，一次扫面数据即可完成聚类 3.可以分离簇和噪声数据	1.不很好反应高尺寸数据 2.不产生完全聚类 3.不很好反应数据集以变化的密度	自动确定簇的数目
EM	4	N	简单稳定	容易陷入局部最优	
FarthestFirst	2	N	速度快	精确度相对较低	
FilteredClusterer	1	Y	可以使用不同的聚类算法与Filter进行聚类	复杂	灵活采用各种Filter和聚类算法
HierarchicalClusterer	2	N	比较简单	当在算法开始阶段，若出现聚类错误，那么这种错误将一直会被延续，无法修改	
MakeDensityBasedClusterer	1	Y	1.局部搜索能力强 2.收敛速度快 3.可以选择聚类算法	1.比较复杂 2.准确度较低	
OPTICS	2	N	1.克服参数设置由用户决定的缺点 2.对输入参数不敏感	1.只是对数据对象集中的对象进行排序，输出一个有序的对象列表(cluster-ordering) 2.比较复杂	不显示地产生数据聚类
sIB	5	N	1.具有较低的时间和空间复杂度 2.保证可以得到问题的局部优解	1.随机选取的初始解导致算法容易陷入局部解 2.压缩变量参数需要由用户指定 3.准确度低	
SimpleKMeans	4	N	1.简单 2.精确度高	不能处理非球形簇，不同尺寸和不同密度的簇	
XMeans	9	N	精确度高	相对比较复杂	

其中, “Y”表示用到其它聚类, “N”表示没有用到其他聚类.

4.3 实验步骤

对数据进行预处理后在 weka 平台上进行聚类实验, 主要步骤概括如下, 如图 1 所示.

4.4 实验结果

考虑到基于一卡通消费数据的挖掘分析必须简单易行, 则上节提到的所有方法中 MakeDensityBased Clusterer 和 FilteredClusterer 方法由于需要对包含的多个聚类算法和多个过滤器进行选择, 首先被排除.

使用同一数据集在相同的实验环境下进行实验, 相同的实验参数设置如 4.2 节所述. 选取下面几个指标进行对比, 实验结果如表 2 所示.

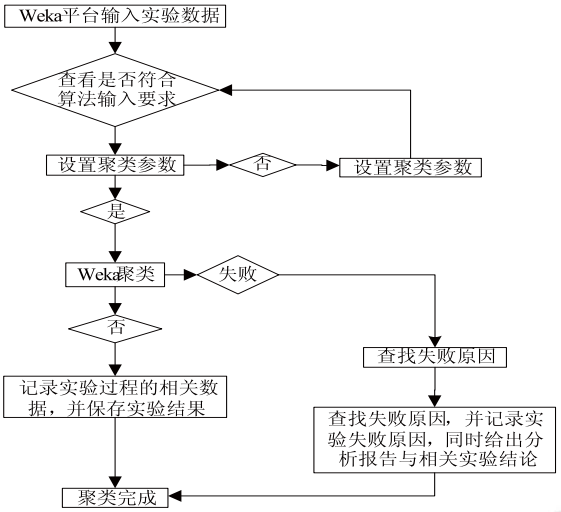

图 1 weka 聚类步骤图

5 实验结果分析

通过上述的实验可以得出，首先从实验的可操作性上来说，HierarchicalClusterer 方法需要选择其他的聚类，并非单纯的聚类算法；而 OPTICS 算法由于复杂性较高，故这两个算法都不适合该数据集上的聚类，难以完成数据的挖掘.

其次，从聚类效果上来讲，Cobweb 算法和 DBScan 算法聚类的效果较差，不能很好地根据数据集实现消费人群的分类，难以为后续的数据分析所使用. 对于该数据集来说这两个都不是很好的聚类方法.

然后，从聚类时间上来看，EM 算法和 sIB 算法挖掘效率较低，时间开销较大. 基于该算法开发相应的数据挖掘系统，系统响应时间过长，不合符高校的应用需求.

另外，从聚类的结果与事实情况相比较来看，FarthestFirst 算法虽然效果较好，但是并未聚类为 4 类，不符合先前已知的分类情况.

K-means 和 XMeans 方法总体来说适合该数据集的数据挖掘.

最后，从简单性上来讲 XMeans 参数过多复杂性较大；从时间上来说，消耗也比 SimpleKMeans 算法大很多.

因此，综合而言，本文认为 K-means 算法最适合做校园一卡通数据的数据聚类挖掘.

表 2 实验结果表

聚类算法	运行时间	实验参数设置	聚类分布	说明
Cobweb	3.53s	acuity:1.0 cutoff: 0.0028209479177387815 seed:42	分类树显示	融合数:420 合并数:295 聚类数:592
DBScan	4.49s	epsilon ： 0.9 minPoint:6	0:1671 (100%)	聚类数:1671 只有1类
EM	10.82s	maxIterations:500 minStdDev: 1.0E-6 numClusters:-1 seed:100	0:396 (24%) 1:490 (29%) 2:375 (22%) 3:410 (25%)	4类
FarthestFirst	0.02s	cluster:SimpleKMeans* filter:AllFilter	0:1276(76%) 1:3 (0%) 2:5 (0%) 3:387 (23%)	4类
HierarchicalClusterer	-	distanceFunction:EuclideanDistance numClusters:4	-	内存溢出
OPTICS	6.3s	Epsilon:0.9 minPoint:0.9	-	数据均未聚类
sIB	11.76s	maxIteration:500 minChange:0 numClusters:4 numRestarts:5 seed:1	0:397 (24%) 1:343 (21%) 2:412 (25%) 3:519 (31%)	4类
SimpleKMeans	0.05s	distanceFunction:EuclideanDistance maxIteration:500 numClusters:4 seed:10	0:134 (8%) 1:656 (39%) 2:407 (24%) 3:474(28%)	4类
XMeans	0.16s	binValue:1.0 cutoffFactor:0.5 distanceFunction:EuclideanDistance maxIteration:500 maxKMeans:1000 maxKMeansForChildren:1000 maxNumClusters:4 minNumClusters:4 seed:10	0:306 (18%) 1:456(27%) 2:374 (22%) 3:535 (32%)	4类

其中，“-”表示没有，*中 SimpleKMeans 的参数设置和下面的 SimpleKMeans 参数一致.

(下转第 183 页)

创建图 3 所示目录树将增大整个方法的时间和空间复杂度, 尤其是在文件数量很大、单个文件较小的情况下. 本节选择了 2 个测试用例来测试该方法的时间和空间复杂度, 测试用例及实验结果见表 1 所示.

表 1 测试用例及结果

序号	文件个数	目录个数	时间(s)	空间(M)
1	53246	6713	55	30.4
2	8216	1563	7	5.0

测试用例 1 的文件和目录个数累计 59959 个对象, 累计大小是 2.33GB, 文件平均大小是 40.7KB, 创建图 3 所示的目录树, 耗时 55 秒, 空间消耗约 30MB. 对于容量为 4.7GB 的 DVD 光盘而言, 如果文件平均大小相近, 那么创建一张 DVD 光盘的目录树, 耗时约 110 秒, 空间消耗约 60MB. 如果文件平均大小增大, 那么性能消耗将降低, 例如测试用例 2, 累计 9779 个对象, 累计大小是 15.3G, 文件平均大小是 1.6MB, 时间和空间消耗明显降低.

4 结语

本文提出了一种改进的光盘文件系统以及数字光盘数据保密的方法, 将光盘的存储空间划分为明文区、保密区和参数区, 将保密区的阅读工具绑定在明文区. 该方法创建的数字光盘兼容所有的光盘驱动器, 与设备无关, 无需安装额外的工具软件, 使用简便, 保密成本低. 如何安全、便捷地打开保密区涉密文件是本文的后续研究工作.

参考文献

1 Tanenbaum AS. Modern Operating Systems. 3rd., New jersey: Prentice Hall, 2009. 175−178.

2 白兆华.基于 UDF 文件系统的蓝光媒体操作软件的研究与实现[硕士学位论文].西安:西安电子科技大学,2009.

3 苏斌.DVD 区域码及区域码管理.实用无线电,2000,5: 49−49.

4 张卫民.一种对光盘数据加密的方法及装置.发明专利, 201010154843.9,2010.

5 王年华.一种光盘数据加密方法.发明专利,200910194378.9,2009.

6 杨耀东.光盘授权播放内容加密算法研究[硕士学位论文].武汉:华中科技大学,2011.

7 肖飞,王运琼,李映松,李必谨.基于光盘映像文件的 CD-ROM 数据加密与解密方法.计算机科学,2009,36(5): 299−301.

(上接第 161 页)

6 结语

本文使用常用的聚类算法在校园一卡通的消费数据上进行了实验. 通过对多个指标的分析, 得出了 K-means 算法最合适在当前的数据上做聚类分析的结论. 该结论将对以后进一步挖掘和应用现有的一卡通消费数据具有十分重要的指导意义. 而文献[4]中使用 K-means 算法在一卡通数据上得出的良好效果也从另一个角度验证了结论的正确性.

参考文献

1 张兵兵,王建,张建威,等.数据挖掘在校园一卡通系统中的应用初探.数理医药学杂志,2009,22(5):572−575.

2 罗华群,易国平.校园一卡通数据的挖掘与应用.科技信息, 2010,(1X):41−41.

3 李珊娜.基于校园一卡通平台的数据挖掘应用研究.铁路计算机应用,2010,(6):55−58.

4 徐剑,陈劲舟.数据挖掘在校园一卡通数据的应用与研究.电脑知识与技术,2012,33.

5 陈慧萍,林莉莉,王建东,等.WEKA 数据挖掘平台及其二次开发.计算机工程与应用,2008,44(19):76−79.

6 魏丽.数据挖掘中聚类算法比较研究.电脑知识与技术, 2007,21.

7 Sharma N, Bajpai A, Litoriya MR. Comparison the various clustering algorithms of weka tools. facilities, 2012, 4: 7.

8 Krishna P, Senguttuvan A, Swarna T, Latha D. Clustering on Large Numeric Data Sets Using Hierarchical Approach: Birch. Global Journal of Computer Science and Technology, 2012, 12(12-C).