挑选聚类算法的网格连通图方法①

李翔宇, 王开军, 郭躬德

(福建师范大学 数学与计算机科学学院, 福州 350007)

摘 要: 每一种聚类算法都有其适合处理的特定分布的数据集. 为了给未知分布数据集挑选合适的聚类算法, 提出了一种挑选聚类算法的网格连通图方法 SCGG. SCGG 通过对数据潜在类结构的分析, 若含有环形结构类则选择层次聚类的单连接算法对数据聚类, 否则选择 k-means 算法. 实验显示该方法十分的有效, 能够挑选到合适的聚类算法对数据聚类.

关键词: 网格连通图; 挑选聚类算法

Selection of Clustering Algorithms Based on Grid-Connected Graph

LI Xiang-Yu, WANG Kai-Jun, GUO Gong-De

(School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)

Abstract: Each clustering algorithm has its suitable treatment of specific distribution data set. The SCGG method based on the Grid-Connected Graph is proposed to select suitable clustering algorithm for unknown distribution data set. The SCGG method analyses the potential structure of the data set. If the data set has ring clustering structure, the method select a single hierarchiral clustering algorithm, otherwise it selects the k-means. Experiment results show that SCGG is very efficient and successful.

Key words: grid-connected graph; selection of clustering algorithms

聚类是将数据对象分成类或簇的过程,使同一个类中的对象之间具有很高的相似度, 而不同类的对象高度相异[1]. 目前, 传统的聚类算法大致可以分成: 层次方法、划分方法、基于密度的方法、基于网格的方法和基于模型的方法, 这些算法的聚类过程一般都采用不同的聚类准则. 对给定的数据集, 不同的聚类算法, 或者甚至一种聚类算法使用不同的参数, 一般会产生不同的聚类结果[2]. 每一种聚类算法对具有某种特定分布的数据相对其他算法会有较好的聚类结果. 例如, K-means 算法仅适合于聚类结果为凸形(即类簇为凸形)的数据集[3], 而对于非凸形(如环形) 的数据,文献[4,5]显示其效果较差. 文献[6]中描述层次聚类的单连接算法对具有环形结构类的数据有较好的聚类效果, 与 K-means 相比, 该算法对具有环形结构类的数据聚类效果更好. 如果事先对数据进行分析并获得数据的潜在类结构, 然后依据备选聚类算法对该类结构数据的适用性, 挑选出更适合这种类结构的聚类算法对数据聚类, 那么可以得到较好的数据聚类效果.

为了实现给未知分布数据集挑选聚类效果更好的聚类算法, 本文提出了一种挑选聚类算法的网格连通图方法(Selection of Clustering Algorithms Based on Grid-Conected Graph, SCGG). 这种方法首先通过对给定数据划分二阶网格空间, 然后根据二阶网格空间上的白网格构建网格连通图, 通过连通图的数目可以发现数据集中潜在的环形(包括有较小缺口的环形)结构类, 最终根据数据中是否含有环形结构类这一特征, 从层次聚类的单连接算法和 k-means 算法中挑选具有最好聚类效果的算法对数据聚类.

① 基金项目:国家自然科学基金(61175123);福建省教育厅资助项目(JA09043);高校产学合作重大项目(2010H6007)
收稿时间:2011-12-29;收到修改稿时间:2012-03-14

1 SCGG方法的相关定义

SCGG 方法通过网格连通图的数目发现数据集潜含的环形类结构, 根据给定的数据集构建网格连通图, 此过程中涉及相关名词的定义如下:

定义 1(数据空间划分)d 维数据集所在的数据空间 $D=A_1 \times A_2 \times A_3 \times \cdots \times A_d$, 其中 $A_i=[Min_i, Min_i+(m+1)(Max_i-Min_i)/m)(1\leqslant i\leqslant d)$, 且 Max_i、Min_i 分别是数据集在第 i 维上的最大值与最小值, m 是维分割参数. 注: 设置 m+1 是为了让某一维上有最大值的数据点能够包含在数据空间中.

定义 2(网格单元)按照定义 1 的方式划分 d 维数据空间, 每一维都划分成 m+1 个间隔段, 则该数据空间被分成$(m+1)^d$ 个大小相等且互不相交的网格单元. 网格单元 U 可以表示为:

$$U=\prod_{i=1}^{d}\left[Min_i+\frac{k_j(Max_i-Min_i)}{m}, Min_i+\frac{(k_j+1)(Max_i-Min_i)}{m}\right)$$

其中, $0\leqslant k_j\leqslant m$, $1\leqslant j\leqslant d$.

定义 3(边缘网格单元)数据空间按定义 2 的网格单元划分后, 每一维上都有一个连续区间 A, 在这个区间的两侧各增加一个长度为$(Max_i-Min_i)/m$ 的间隔段, 由此增加了若干的网格单元, 称这些网格单元为边缘网格单元.

定义 4(网格特征向量)u(id, den, mean)代表一个网格单元(包括边缘网格单元)的网格特征向量. id 为网格编号, 表征该网格单元在网格空间中的位置信息; den 为该网格单元的密度; 网格质心点(简称"质心点"):

$$mean=\sum_{q=1}^{den}v_q/den$$

其中, v_q 是落入该网格单元的第 q 个数据点.

定义 5 (黑白网格). u.den>Minpts 时, 称该网格单元为黑网格, 否则称为白网格, 其中 Minpts 为该网格单元所处的网格空间上的密度阈值.

定义 6 (一阶网格空间). 按照定义 2 的网格划分方式将 d 维数据集的数据空间划分得到的网格空间称为一阶网格空间.

定义 7 (二阶网格空间). 首先在一阶网格空间中, 黑网格的质心点构成一个质心数据集; 然后将质心数据集划分成定义 1 中描述的数据空间; 最后分割该数据空间, 使其被分割成定义 2 的网格单元, 同时按照定义 3 描述的方式增加边缘网格单元. 由以上步骤得到的网格空间称为二阶网格空间.

定义 8 (网格单元连通). 网格空间的两个网格单元, 它们之间存在公共边时认为这两个网格单元连通.

定义 9 (网格连通图). 以网格单元作为为图的节点, 从每个节点出发, 以边(两个网格单元是网格连通时, 认为两个节点之间有边)为路径, 可以到达图中的任意节点.

2 SCGG方法

SCGG 方法可以分成数据类结构分析与聚类算法选择两个部分. 数据类结构分析分为构建二阶网格空间与构建网格连通图. 对给定数据集, SCGG 方法首先通过构建二阶网格空间; 然后根据该空间上的白网格单元构建网格连通图, 并依据网格连通图的数目判断数据集中是否潜含环形结构的类; 若含有环形结构的类, 则挑选单连接算法聚类, 否则, 选用 k-means 聚类.

2.1 二阶网格空间的构建

首先将给定数据集的数据空间划分成一阶网格空间, 然后依据定义 6 将数据点映射到一阶网格空间中, 这样可以将数据集的类结构问题转换成该空间上黑网格的结构问题.

具有环形结构类的数据投射到一阶网格空间后, 可能会出现图 1 中所示的环形结构(黑网格在其所在空间上构成的结构, 使得一部分白网格与空间上其他白网格无法连通时, 我们认为黑网格构成环形结构)有缺口. 提取一阶网格空间中黑网格的质心点可构成由质心点组成的数据空间, 将质心点数据空间划分网格空间(该网格空间中的网格单元粒度相对变大), 这样就可以使得类似图 1 中有缺口的环形结构转换成图 2 的环形结构.

质心点数据空间划分成二阶网格空间时, 首先将质心点数据空间按照定义 2 划分网格单元, 然后增加边缘网格单元构成二阶网格空间. 增加边缘网格单元的目的是为了避免黑网格在网格空间维度的两端时造成白网格不能连通. 如图 2 所示, 1、3、7、9 号白网格分别单独构成一个网格连通图, 而 SCGG 方法的目的是让 1、3、7、9 这几个网格单元在同一个网格连通图中. 增加边缘网格单元后可以使得这些网格能够连通成一个网格连通图, 如图 3 所示, 13 号白网格构成一个连通图, 其他白网格构成一个网格连通图.

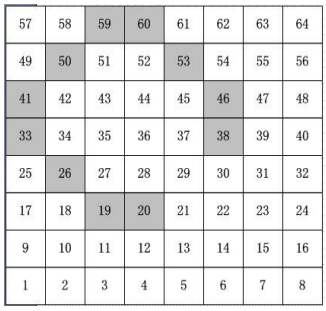


图 1 数据点划分空间

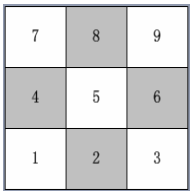


图 2 质心点划分空间

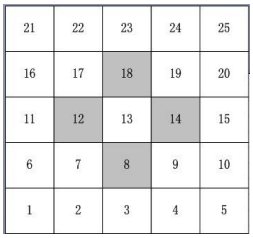


图 3 增加边缘网格后的空间

二阶网格空间的构建步骤如下:

1)根据给定的数据计算网格分割参数 m, 并以此构建该数据集的数据空间. 然后对该数据空间每一维的区间划分出 m+1 个间隔段, 由此可以分割出$(m+1)^d$的网格单元, 这些网格单元构成了一阶网格空间;

2)分配数据点到由步骤 1 所得到的一阶网格空间中, 然后计算该网格空间中各个网格单元的密度, 并依据定义 5 由一阶网格空间的密度阈值 Minpts 区分黑、白网格, 最后计算黑网格质心点;

3)由步骤 2 中的黑网格质心点建立质心点数据集. 类似于步骤 1 的方式将质心点数据集的数据空间划分成若干个网格单元, 这些网格单元在同一维上的间隔段构成的一个连续的区间, 并在这个区间的两侧各增加一个该维的间隔段, 这样就增加了一定数量的网格单元(即边缘网格单元). 由划分质心点数据集的数据空间得到的网格单元与虚拟网格单元构成一个二阶网格空间. 最后分配质心点数据到二阶网格空间中.

2.2 网格连通图的构建

数据集的潜含类结构的分析问题通过二阶网格空间转换成了黑网格在网格空间中的分布结构的分析问题, 而网格空间中白网格是黑网格未占据的空间, 通过分析白网格在网格空间的结构可以对应地体现黑网格的结构. 因此, 我们从分析白网格在网格空间中的分布结构这一角度开展工作. 首先对二维数据集构建二阶网格空间, 在该网格空间上构建白网格连通图, 然后分析白网格的分布结构, 根据白网格连通图的数目判断数据中是否含有环形结构类. 环形结构类将一些白网格封闭在环形内, 与其它白网格区域相分离, 使得白网格区域多于 1; 若无环形结构类, 则所有的白网格都连通在一起.

网格连通图的具体构建: 给定一个包含网格单元的网格集(即节点集), 以及一个由这些节点的边(两个网格单元是连通时, 认为这两个网格所代表的节点之间有边)构成的边集, 构建定义 9 所描述的网格连通图. 二维数据构建二阶网格空间过程中, 构建网格连通图的步骤设计如下:

1)设置当前类标号 k=1, 且设置节点集里所有节点的类标号为 0;

2)不回放地从节点集里任选一个节点, 将其存入网格连通图集;

3)从网格连通图集里任意取出一个类标号为 0 的节点, 首先标记该节点类标号为当前类标号 k, 然后在二维网格空间中按照上、左、下、右的顺序搜索与该节点存在边的节点. 判断搜索得到的节点是否包含于节点集, 若是, 从节点集里取出搜索所得的节点, 将其放入网格连通图集, 否则不处理;

4)重复步骤 3, 直到网格连通图集所有节点的类标号都不为 0;

5)若节点集为空则算法结束, 否则令 k=k+1(寻找下一个网格连通图), 转步骤 2.

2.3 SCGG 方法的实现

SCGG 方法的实现步骤如下:

1)给定一个 d 维数据集, 根据数据点数目 n、各维度上的最大值 Max_i、最小值 Min_i(其中 $1 \leqslant i \leqslant d$), 计算维分割参数 m, 根据定义 1 的方式划分成有界的 d 维数据空间 D.

2)将步骤 1 的数据空间 D 按照 2.1 的步骤组建二阶网格空间. 为所有网格单元的网格特征向量赋值(id: 按照从 1 开始的顺序为二维二阶网格空间的网格单元标注 id 号, 如图 3 中所示, id=8 代表在这个二维二阶网格空间中的第 2 行, 第 3 列; den: 该网格单元分配所得的一阶网格空间黑网格质心点的个数; mean 设置为 0).

3)对于步骤 2 的所有网格单元, 根据二阶网格空间上的密度阈值 Minpts 将它们分成黑、白网格(当网格特征向量的 den 大于 Mintpts 时, 为黑网格, 否则为白网格), 然后以所有白网格建立一个网格集合, 该集合

按照 2.2 的步骤构建网格连通图，最终生成了一个网格连通图集.

4)在网格连通图集里，若节点类标号相同的网格的数目只有 1 个(由 1 个白网格构成的结构无法表征数据环形结构类的特点)时，则在网格连通图集里删除这些网格；计算网格连通图集里不同节点类标号的数目(即白网格连通图的个数)；由连通图的数目就可以判断二维数据中是否含有环形结构类. 连通图的数目大于 1 时(即环形结构类将一些白网格封闭在环形内，与其它白网格区域相分离，使得白网格区域多于 1)，存在环形结构类，否则不存在.

5)如果数据中含有环形结构类，选用层次聚类的单连接算法对数据聚类，否则挑选 K-means 算法聚类.

2.4 参数的选取

2.4.1 维分割参数 m 的选取

类似于文献[7,8]中参数设置，维分割参数 m 设置为 $m=\sqrt{n}$，在一阶网格空间中 n 代表了给定的数据点的个数，在二阶网格空间中则代表了质心点数据集的质心点个数.

2.4.2 密度阈值 Minpts 的选取

对给定的数据集，在一阶网格空间中，当密度阈值 Minpts 取大于 0 的较小整数(例如图 1)时，将会滤除稀疏数据点，这样在一阶网格空间中能寻找到较密集的数据点构成的类结构. 随着 Minpts 增大，可得到更为密集的数据点构成的类结构，但这样所得的类结构与原始数据点构成的类结构可能有较大的偏差. 如果 Minpts=0，则是对原始数据集在一阶网格空间中寻找类结构. 二阶网格空间是由一阶网格空间中的黑网格质心点集构建，将不对二阶网格空间进行数据点的滤除工作，故在二阶网格空间上设置密度阈值 Minpts=0.

3 实验分析

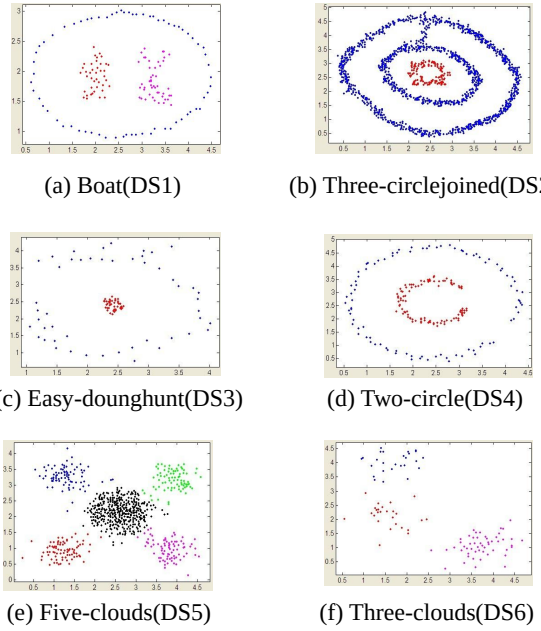
实验中 SCGG 方法采用 Matlab 语言编写，其中参数设置为：网格参数 m=$\sqrt{n}$ (n 为数据点的个数；对二阶网格空间，n 为一阶网格空间黑网格质心点的个数)，密度阈值 Minpts=0 . 实验中 K-means 算法采用 CVAP 工具[9]的程序，运行程序 10 次取均方误差最小的结果作为 K-means 算法的结果. 层次聚类的单连接算法程序采用 Matlab 软件自带的算法程序. 实验中的算法均使用欧式距离作为样本之间距离测度.

本实验采用 6 组二维人工数据集[10-13](如图 4)进行测试，其中 DS3, DS4 均为删除文献[10]、[13]的数据集中部分数据点而得到的带较小缺口的环形结构类. 这 6 组数据用于 SCGG 方法的有效性测试(由于没有类似的方法可以与 SCGG 方法进行性能的比较，本实验将检验 SCGG 方法是否能够挑选出正确的聚类算法对数据聚类). 这 6 组数据的具体描述如表 1 所示.

表 1　二维人工数据集的特征

数据集名称	数据对象个数	类的个数	环形类的个数
DS1	157	3	1
DS2	1354	2	1
DS3	90	2	1(有缺口)
DS4	191	2	2(都有缺口)
DS5	690	5	0
DS6	104	3	0

实验选用数据集的数据点的数值每一维均在区间(0, 5). 这些数据集如图 4 所示:



(a) Boat(DS1)　(b) Three-circlejoined(DS2)

(c) Easy-dounghunt(DS3)　(d) Two-circle(DS4)

(e) Five-clouds(DS5)　(f) Three-clouds(DS6)

图 4　6 组人工数据集

为了检测 SCGG 方法的有效性，对表 1 中的 6 组数据进行测试，考虑没有类似的方法可以与 SCGG 方法进行性能的比较，故设计实验为：只选用 K-means 算法、只选用层次聚类的单连接算法对数据聚类，以及 SCGG 方法挑选适合该数据集的算法对数据聚类. 实验结果如表 2 所示.

由表 2 可知，对于前 4 组具有环形结构类(包括 DS3 和 DS4 这两组有缺口环形结构类)的数据集，本文

的算法可以准确的选择单连接算法对其聚类, 对于后2组不存在环形类结果的数据集, 算法挑选了K-means算法对其聚类. 故通过实验可知, SCGG 方法可以准确的选择合适于数据集的聚类算法对其聚类.

表 2 K-means 算法、单连接算法和 SCGG 方法对数据集聚类结果的错误率

数据集	DS1	DS2	DS3	DS4	DS5	DS6
K-means	38.8535%	50.4431%	30%	51.3089%	1.0245%	0%
单连接	0%	0%	0%	0%	45.942%	24.0385%
SCGG	0%	0%	0%	0%	1.0245%	0%
SCGG挑选出的算法	单连接	单连接	单连接	单连接	K-means	K-means

4 结论与展望

SCGG 方法通过构建二阶网格空间将未知分布数据转换成黑网格在二阶网格空间的分布结构信息. 通过二阶网格空间的白网格构建网格连通图间接地体现黑网格的分布结构; 再由白网格连通图的数目判断给定数据集是否含有环形结构类, 若有则挑选单连接算法对数据聚类, 否则挑选k-means算法. 实验显示该算法能够正确发现环形结构类(包括有较小缺口的环形类), 并用最佳的聚类算法对数据进行聚类. 本文只针对两种聚类算法的挑选问题进行了研究, 下一步将对更多聚类算法的挑选开展工作.

参考文献

1 Han JW, Kambr M. 数据挖掘概念与技术.第 2 版.北京:机械工业出版社,2001.251−305.

2 Jiang DX, Tang C, Zhang AD. Cluster analysis for gene expression data: A survey. IEEE Trans. on Knowledge and Data Engineering, 2004,16(11):1370−1386.

3 孙吉贵,刘杰,赵连宇.聚类算法研究.软件学报,2008,19(1): 48−61.

4 孙雪,李昆仑,胡夕坤,赵瑞.基于半监督 K-means 的 K 值全局寻优算法.北京交通大学学报,2009,33(6):106−109.

5 潘晓英,刘芳,焦李成.密度敏感的多智能体进化聚类算法.软件学报,2010,21(10):2420−2431.

6 范明.聚类算法在 web 挖掘中的应用.西安:西北工业大学, 2007.

7 王建会,申展,胡运发.一种实用高效的聚类算法.软件学报,2004,15(5):697−705.

8 周炎涛,吴正国,易兴东.基于网格带有参考参数的扩展聚类算法.湖南大学学报(自然科学版),2009,36(2):48−52.

9 Wang KJ, Wang BJ, Peng LQ. CVAP: Validation for Cluster Analyses. Data Science Journal,2009,8(20):88−93.

10 Kuncheva LI, Vetrov DP. Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006,28(11):1798−1808.

11 Ng AY, Jordan MI, Weiss Y. On Spectral Clustering:Analysis and an Algorithm. Proc. of 14th Advances in Neural Information Processing Systems. 2001 849−856.

12 Lange T, Roth V, Braun ML, Buhmann JM. Stability-Based Validation of Clustering Solutions. Neural Computation, 2004,16(6):1299−1323.

13 公茂果,王爽,马萌,曹宇,焦李成.马文萍.复杂分布数据的二阶段聚类算法.软件学报,2011,22(11):2760−2772.