改进型规模约束在聚类算法中的应用

朱顺痣[1], 符长虹[2], 刘利钊[1], 洪文兴[2]
(1 厦门理工学院 计算机科学与技术系, 福建 厦门 361024; 2 厦门大学 自动化系, 福建 厦门 361005)

摘 要: 规模约束可有效改善聚类算法的性能, 但是各类规模约束后所含实例对象数量不一致将降低聚类算法的性能. 采用一种新的模式对各类进行了规模约束, 并转化为线性规划问题进行求解. UCI 标准数据集上的实验结果表明本算法与随机模式相比具有更好的聚类精度, 即使当规模约束适当放宽后, 聚类性能也可得到明显提升. 提出的方法能够有效地提高聚类的准确性.

关键词: 规模约束; 聚类; 线性规划; 随机模式
中图分类号: U666.72 文献标识码: A 文章编号: 1000-7180(2011)08-0169-04

Application of Improved Size Constrains in Clustering Methods

ZHU Shun-zhi[1], FU Chang-hong[2], LIU Li-zhao[1], HONG Wen-xing[2]
(1 Department of Computer Science and Technology, Xiamen University of Technology, Xiamen 361024, China;
2 Department of Automation, Xiamen University, Xiamen 361005, China)

Abstract: Size constraints can improve the clustering performance of clustering methods. However the differences in the size of clusters, i.e. the number of instances contained in each cluster will decrease the clustering performance. This paper introduces a new scheme of size constraints on size of each cluster and transforms them into linear programming optimization. Experiments results on UCI benchmark datasets show that the new method outperforms the random scheme. The clustering performance can be increased even when the size constraints are relaxed to some extent. The new algorithm can increase the clustering accuracy efficiently.

Key words: size constraints; clustering; linear programming; random mode

1 引言

实际应用中, 在执行聚类之前, 通常可以获取样本关系之间的背景信息数据或者各组类的大致规模, 这些先验信息原本对聚类工作有很大帮助, 但是传统的聚类算法并不能有效利用这些先验信息来更好地优化聚类. 为此国内外研究者考虑利用实例背景信息, 如样本两两之间的 must-link 或者 cannot-link 作为约束条件: 若两样本处在同一组类中, 我们称之为 must-link, 否则, 称之为 cannot-link. Wagstaff 等人[1]在满足各组类相互关系的约束下, 将这种形式的背景信息应用于 K 均值聚类算法中. Zhu[2] 和 Basu 等[3] 也在聚类研究中考虑两两样本约束及样本的基本指标. 国内研究者也在此方面做了一些工作[4]. 文献[5] 在研究约束聚类的距离测度做了相关工作, 当组类具有相似的规模或重要性时, 还有一些针对平衡约束的相关工作. 除了某些特定的实际应用要求外, 平衡约束对生成更多有意义的初始组类以及避免不合适的组类有很大的帮助. Baerjee 等[6] 表明在满足平衡约束的条件下, 少量样本就能够获取组类质心并在其他样本分类时归入其中. Zhang 等[7] 也对平衡约束进行相关讨论, 并为样本分类提出一种遍历二分启发式方法. 以上所有相关工作都可以证明利用背景先验知识可以增加聚类

收稿日期: 2011-05-16; 修回日期: 2011-06-20
基金项目: 国家自然科学基金项目(61070151); 福建省自然科学基金项目(2010J01353); 福建省仿脑智能系统重点实验室(厦门大学)开放基金项目(BLISSOS2010102)

表达的准确性和可扩展性[8-10].

本文将平衡约束扩展到规模约束, 基于数据分布的先验知识, 首先分配每个组类的规模, 然后试图找到一个满足规模约束的划分, 接着同时考虑规模约束和实例 cannot-link 约束, 并提出一种启发式算法模型, 通过将其转化为整型线性规划最优模型来解决此约束聚类问题. 基于 UCI 标准数据库的实验表明, 本文提出的方法能够有效地提高聚类的准确性, 为聚类提供了新思路与方法.

2 算法原理

在规模约束聚类问题中, 各组类样本数量为先验知识, 并且可以通过任意一种传统聚类方法(如均值法) 得到样本划分结果. 接下来, 问题提出如下:

给出含有 n 个样本的一组数据, 令 $\boldsymbol{A} = (A_1, A_2, \cdots, A_p)$ 为样本的 P 个组类划分, 并且令 $NumA = (na_1, na_2, \cdots, na_p)$ 为各组类的样本数量, 在此, 寻找样本的另一划分 $\boldsymbol{B} = (B_1, B_2, \cdots, B_p)$ 使得划分 A、B 的一致性达到最大, 然后令 $NumB = (nb_1, nb_2, \cdots, nb_p)$ 为各组类的规模约束, 其中当 $nb_1 = nb_2 = \cdots = nb_p$ 时, 被称为 1∶1 模式. $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 能可表达成一个 $n \times p$ 的划分矩阵. 矩阵的每一行表示一个样本, 每一列表示一个划分. 在划分 $\boldsymbol{A}$ 或划分 $\boldsymbol{B}$ 中, 当样本 i 属于组类 j 时, 则记作 $a_{ij} = 1$ 或者 $b_{ij} = 1$.

举例如下:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ & & & \cdots & & & \\ & & & \cdots & & & \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ & & & \cdots & & & \\ & & & \cdots & & & \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$$

其中

$$\sum_{i=1}^{n} a_{ij} = na_j, j = 1, \cdots, p$$

$$\sum_{j=1}^{p} a_{ij} = 1, \ i = 1, \cdots, n$$

由上易知, $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ 是一个 $n \times n$ 阶矩阵, 取值如下:

$$(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})_{ij} = \begin{cases} 1, \text{当 } i、j \text{ 属于划分} \boldsymbol{A} \text{ 中的同一组类时} \\ 0, \text{否则} \end{cases}$$

现在我们需要寻找另一种划分 $\boldsymbol{B}$, 使 $\| \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} \|$ 最小, 即

$$\mathrm{Minimize} \| \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} \| \quad (1)$$

其中

$$\sum_{i=1}^{n} b_{ij} = nb_j, j = 1, \cdots, p$$

$$\sum_{j=1}^{p} b_{ij} = 1, \ i = 1, \cdots, n$$

此过程类似于寻找一个划分使得它与一个已知划分达到一致性的最大化.

3 规模约束聚类的求解

本文结合整型线性规划, 将规模约束聚类问题转化为一个最优化问题, 提出一种启发式算法更简单、快捷来求解规模约束聚类.

3.1 启发式细节过程

为了解决前一节所叙述的问题, 首先, 本文定义

$$\boldsymbol{D}_a = \mathrm{diag}(na_1, na_2, \cdots, na_p)$$

以及

$$\boldsymbol{D}_b = \mathrm{diag}(nb_1, nb_2, \cdots, nb_p)$$

令

$$\boldsymbol{U}_j = \frac{1}{\sqrt{na_j}} \begin{bmatrix} a_{1j} \\ a_{2j} \\ \cdots \\ \cdots \\ \cdots \\ a_{nj} \end{bmatrix}, j = 1, \cdots, p$$

式中, $a_{ij} \in \{0, 1\}$.

且有

$$\sum_{i=1}^{n} a_{ij} = na_j, j = 1, \cdots, p,$$

$$\sum_{j=1}^{p} a_{ij} = 1, i = 1, \cdots, n.$$

由此, 易知 $\boldsymbol{U} = \boldsymbol{A}(\boldsymbol{D}_a)^{-1/2}$ (2)

同理,

令

$$\boldsymbol{V}_j = \frac{1}{\sqrt{nb_j}} \begin{bmatrix} b_{1j} \\ b_{2j} \\ \cdots \\ \cdots \\ \cdots \\ b_{nj} \end{bmatrix}, j = 1, \cdots, p$$

式中, $b_{ij} \in \{0, 1\}$

且有

$$\sum_{i=1}^{n} b_{ij} = nb_j, j = 1, \cdots, p,$$

$$\sum_{j=1}^{p} b_{ij} = 1, i = 1, \cdots, n.$$

由此, 易知 $\boldsymbol{V} = \boldsymbol{B}(\boldsymbol{D}_b)^{-1/2}$　　(3)

综合式(2) 与式(3), 易得

$$\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{D}_a\boldsymbol{U}^{\mathrm{T}}, \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}} = \boldsymbol{V}\boldsymbol{D}_b\boldsymbol{V}^{\mathrm{T}} \quad (4)$$

结合式(1), 故有

$$\begin{aligned} \|\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\|^2 &= \|\boldsymbol{U}\boldsymbol{D}_a\boldsymbol{U}^{\mathrm{T}} - \boldsymbol{V}\boldsymbol{D}_b\boldsymbol{V}^{\mathrm{T}}\|^2 \\ &= \|\boldsymbol{D}_a - \boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}_b(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V})^{\mathrm{T}}\|^2 \end{aligned} \quad (5)$$

若存在一个 $\boldsymbol{V}$, 使得

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{J} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ & & & \cdots & & & \\ & & & \cdots & & & \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

那么, 易得

$$\|\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\|^2 = \|\boldsymbol{D}_a - \boldsymbol{D}_b\|^2 \quad (6)$$

其中, $\boldsymbol{B}$ 为划分 $\boldsymbol{A}$ 的最优近似解, 且满足规模约束. 但是, 在实际问题中, 很难找到这样的 $\boldsymbol{V}$ 满足 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{J}$. 因此, 本文将选择一个 $\boldsymbol{V}$ 使得 $\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V} - \boldsymbol{J}\|$ 最小.

$$\|\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V} - \boldsymbol{J}\| = 2p - 2\sum_{i=1}^{n}\sum_{j=1}^{p} u_{ij}v_{ij} \quad (7)$$

式中, u_{ij}, v_{ij} 分别为向量 $\boldsymbol{U}_j$ 和 $\boldsymbol{V}_j$ 的元素. 因此, 此规模约束的聚类问题可以转化为如下线性规划问题:

$$\text{Minimize} - \left[\sum_{j=1}^{p} \frac{1}{\sqrt{na_j nb_j}} \sum_{i=1}^{n} a_{ij}b_{ij}\right] \quad (8)$$

式中,

$$\sum_{i=1}^{n} b_{ij} = nb_j, j = 1, \cdots, p,$$

$$\sum_{j=1}^{p} b_{ij} = 1, i = 1, \cdots, n,$$

且有 $b_{ij} \in \{0, 1\}$.

式(8) 是一个典型的二元整型线性规划问题, 可以很容易地用任何线性规划算法解决.

3.2　启发式过程总结与关键步骤

综上所述, 启发式过程有两个步骤:

(1) 采用高效和有效的传统或实例约束聚类算法处理样本数据;

(2) 在样本数据的先验知识基础上, 创建规模约束, 接着使用本节介绍的方法将规模约束聚类转化为二元整线性规划问题.

还有一个值得注意的事情是: 当我们使用任何其他实例约束聚类算法作为我们的基本算法时, 一些约束可能与最终结果相违背. 实际上, 如 cannot-link 约束可以被合并到本文已转化的线性规划模型中成为一个不等式约束模型. 例如, 若样本 k 和样本 l 处在不同组类时, 约束条件可以表示为

$$b_{kj} + b_{lj} \leqslant 1, j = 1, \cdots, p$$

因此, 本文提出的方法将先验规模信息加入实例约束聚类算法中来优化模型或者直接使用传统的聚类方法解决约束型聚类问题.

表 1　采用小数据原则、大数据原则与第 1 种方法比较的实验结果

数据集类型	算法	准确度	平均信息量	ARI	NMI
Iris	第 1 种方法	0.846 7	0.255 9	0.641 6	0.675 0
	小数据原则	0.846 7	0.255 9	0.641 6	0.675 0
	大数据原则	0.846 7	0.255 9	0.641 6	0.675 0
Wine	第 1 种方法	0.707 9	0.440 0	0.386 3	0.438 3
	小数据原则	0.816 5	0.327 5	0.491 9	0.545 6
	大数据原则	0.853 4	0.301 3	0.524 5	0.572 4
Balance Scale	第 1 种方法	0.524 8	0.702 0	0.138 9	0.093 1
	小数据原则	0.637 6	0.581 7	0.242 1	0.182 4
	大数据原则	0.661 7	0.538 6	0.289 8	0.210 3
Ionosphere	第 1 种方法	0.803 4	0.353 4	0.405 6	0.292 6
	小数据原则	0.894 9	0.269 8	0.515 3	0.384 7
	大数据原则	0.914 7	0.221 3	0.564 4	0.425 2

4 实验结果对比分析

表 1 说明了本文使用小数据原则与大数据原则方法进行启发式规模约束聚类所得到的评价结果. 由表中可以清楚地看到, 本文采用小数据原则与大数据原则进行规模约束, 聚类性能得到了提高, 得到了优化, 将小数据原则与大数据原则规模约束条件加入到传统的或现有的随机规模约束的聚类算法中, 仍然无需确定精确的聚类规模, UCI 数据集上的实验结果表明聚类性能有了显著提升.

使用 UCI 中四个数据集并且通过放宽精确规模约束到范围规模约束来研究本文所提方法的灵活性, 这意味着除了指定特定的规模给各个组类, 大致的规模范围同样可以在本文所提方法中进行应用, 这在现实应用中非常重要. 本文指定各组类的样本数量不超过 75, 对数据采用 1:1 模式与随机模式进行比较结果如表 2 所示.

表 2 规模约束放宽后的案例研究结果

数据集类型	算法	准确度	平均信息量	ARI	NMI
Wine	规模约束放宽	0.846 7	0.255 9	0.641 6	0.675 0
	1:1 模式放宽	0.930 1	0.198 9	0.763 7	0.792 8

5 结束语

本文针对规模约束环节来进行改进, 以强化聚类算法的性能并以此来优化数据挖掘的过程: 对各类规模约束后所含实例对象数量不一致将降低聚类算法的性能的具体问题, 采用一种新的模式对各类进行了规模约束, 并将这个问题转化为了线性优化问题, 对线性优化问题进行了最优处理和求解, 从而改进规模约束问题的解; 通过 UCI 标准数据集上的实验, 说明这个算法与随机模式相比具有更好的聚类精度, 也就是说使当规模约束适当放宽后, 聚类性能也可得到明显提升.

参考文献:

[1] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge[C]// Proceedings of ICML. San Francisco, CA, USA: ACM, 2001: 167-173.

[2] Zhu Shunzhi, Wang Dingding, Li Tao. Data clustering with size constrains[J]. Knowledge-Based Systems, 2010: 883-889.

[3] Basu S, Banerjee A, Mooney R J. Active semi-supervision for pair-wise constrained clustering [C]// Proceedings of SIAM Data Mining. Florida, USA: ACM, 2004: 111-119.

[4] 钟洪, 夏利民. 基于互信息约束聚类的图像语义标注[J]. 中国图形图像学报, 2009, 14(6): 25-28.

[5] 李敏强, 李智. 基于约束聚类的一种概念学习方法[J]. 系统工程学报, 2004, 19(5): 32-36.

[6] Banerjee A, Ghosh J. Scalable clustering algorithms with balancing constraints [J]. Data Mining Knowledge Discovery, 2006: 31-34.

[7] Zhong S, Ghosh J. A Unied framework for model-based clustering [J]. Journal of Machine Learning Research, 2003: 12-15.

[8] Massatfa H. An algorithm to maximize the agreement between partitions [J]. Journal of Classification, 1992: 22-24.

[9] Hubert L, Arabie P. Comparing partitions [J]. Journal of Classification, 1985.

[10] Studholme C, Hill D, Hawkes D J. An overlap invariant entropy measure of 3D medical image alignment [J]. Pattern Recognition, 1999.

作者简介:

朱顺痣 男, (1973-), 博士, 副教授. 研究方向为信息系统、数据挖掘、GIS 应用研究.

洪文兴 男, (1981-), 博士, 助理教授. 研究方向为推荐系统与数据挖掘.

(上接第 168 页)

[4] Wang R Y. A product perspective on total data quality management[J]. Communications of the ACM, 1998, 41(2): 58-65.

[5] 陈伟, 王昊, 朱文明. 一种提高相似重复记录检测精度的方法[J]. 计算机应用与软件, 2006, 23(10): 29-30.

[6] 陈伟, 丁秋林. 数据清理中不完整数据的清理方法[J]. 微型机与应用, 2005, (2): 44-45.

作者简介:

汤怀美 女, (1984-), 硕士研究生. 研究方向为信息资源管理.