拉普拉斯加权聚类算法

黄鹏飞,张道强
(南京航空航天大学计算机科学与工程系,江苏南京 210016)

摘 要: 本文提出了一种用于聚类分析的加权聚类算法,通过利用拉普拉斯权,将聚类对象之间的结构信息自动转换为对象的权重.由于拉普拉斯权能够描述数据的邻域结构,从而能够更好的聚类.该加权聚类算法在性能上比经典聚类算法有较大改进,还具有对孤立点鲁棒、适合类别不平衡数据聚类、对聚类个数不敏感等优点.人工数据集以及 UCI 标准数据集上的实验证实了本文算法的可行性和有效性.

关键词: 聚类; 拉普拉斯; 加权
中图分类号: TP391.4 **文献标识码**: A **文章编号**: 0372-2112 (2008) 12A-050-05

Weighted Laplacian Clustering Algorithm

HUANG Peng-fei, ZHANG Dao-qiang
(*Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics. Nanjing, Jiangsu* 210016, *China*)

Abstract: In this paper, we propose a novel weighted clustering algorithm based on Laplacian weight, which can automatically transform the structure information between clustering objects into weights of objects. Because Laplacian weight can indicate the neighborhood structure of original data set, better clustering is achieved. Performed on conventional C-means or fuzzy C-means methods, the proposed Laplacian weighting scheme can effectively improve the clustering performance. In addition, the new algorithm achieves some extra advantages such as robustness to outliers, suitability for class-imbalance data clustering and insensitivity to number of clusters, etc. Experimental results on artificial datasets and UCI machine learning repository validate the effectiveness of the proposed algorithm.

Key words: clustering; Laplacian; weighted

1 引言

聚类是将未知类别标号的样本集划分为内在的多个类别,并使得同一个类内的样本具有较高的相似度;而不同类中的样本具有较低的相似度.主要的聚类分析方法可分为如下几类:划分方法,层次聚类方法,基于密度的方法,基于网格的方法和基于模型的方法.本文是主要针对划分方法的研究.

传统的划分聚类方法是从硬划分开始的,算法聚类时是把每一个对象严格的划分到某一个聚类中,要求非此即彼,即每一个对象在且只能在一个聚类中,这样的划分使得类别界限十分分明.C-均值(HCM)是此类算法的典型代表.然而,在客观世界中,大多数对象并没有严格的属性,它们在状态和类属方面存在着中介性,因而进行软划分可能会更加的真实和合理,由此出现了各种软划分算法.此类算法最典型的代表便是模糊 C-均值(FCM).HCM 和 FCM 已经得到了非常广泛的应用,并且出现了很多变种,如联机版本 HCM[1],选择性 HCM[2],运用模糊协方差矩阵的模糊聚类[3]等.

HCM 和 FCM 均有缺点和不足,如 HCM 准则函数不可微等,虽然 FCM 引入样本到中心的隶属度,解决了准则函数不可微的不足,但依然无法克服采用平方误差和准则使得方法仅适用于发现球形和类似球形分布的类别等缺点.许多学者在 FCM 的基础上,通过修改准则函数从而达到对不同形状分布样本的聚类.如文献[2,4]中讨论的另外一种距离度量准则,适合环形分布样本聚类的模糊 C-球壳算法[5].除此之外,R.O Duda 和 P.E Hart[6]还提出用类内散度矩阵的行列式作为准则函数,K.L Wu 和 M.S Yang[2]提出一种指数型准则函数,得到的算法可以有效发现大小差别很大的类别,并且对噪声不敏感.

加权聚类算法是一种最近才引起人们注意的算法,能够有效地提高聚类性能.针对样本矢量中各维特征对分类的不同影响,J.Z Huang[7]等人提出了自动变量加

收稿日期:2008-05-07;修回日期:2008-08-19
基金项目:国家自然科学基金(No.60505004);江苏省自然科学基金(No.BK2006521)

权 C 均值算法能够合理的选择变量及每个变量的贡献大小.在传统的基于划分的聚类方法中,一般都是对聚类样本同等对待,但实际上,不同的样本对聚类结果有不同的贡献. C. Z Zhang[8]等人提出基于样本加权的文本聚类算法,利用论文之间的引用关系计算每篇论文的 PageRank 值,并将其作为权重,实现对不同样本的影响差异赋予不同权重,得到更加合理的聚类中心,但该算法仅适用于文本聚类及相关领域. X. B Gao 等人[9]提出一种基于加权模糊 C 均值聚类与统计检验指导的多阈值图像自动分割算法,利用灰度直方图将各灰度级出现的概率作为加权系数,从而提高算法分割效率,但算法主要应用于图像数据. 本文则应用拉普拉斯加权[10~12]调整了 HCM 和 FCM 算法的目标函数,提出了拉普拉斯加权聚类算法. 通过计算样本与聚类中心的距离获得合理的权系数,能够很好地描述数据邻域结构,并能够利用不同样本影响的差异性,自动调整权系数,提高了聚类的性能,且适用于普通样本数据. 文中算法可以有效地反映数据集的邻域结构,从而有很好的鲁棒性,且适合类别不平衡数据聚类. 在人工数据集和 UCI 标准数据集[13]上验证了本文所提算法的有效性.

2　C-均值与模糊 C-均值

2.1　C-均值算法(HCM)

HCM 核心思想如下:算法把 n 个向量 $x_j(j=1,2,\cdots,n)$ 分为 c 个聚类 $G_i(i=1,2,\cdots,c)$,计算每个聚类中心 v_i,使得目标函数值最小.

HCM 的目标函数为:

$$J=\sum_{i=1}^{c}\sum_{x_j\in G_i}\| x_j - v_i \|^2 \tag{1}$$

2.2　模糊 C-均值算法(FCM)

FCM 用隶属度确定每个样本属于某个聚类的程度. 作为 HCM 算法的推广,FCM 把 n 个向量 $x_j(j=1,2,\cdots,n)$ 分为 c 个模糊聚类,并计算每个聚类中心,使得目标函数值最小. FCM 与 HCM 的主要区别在于 FCM 用模糊划分,通过隶属度函数来确定每个样本属于各个聚类的程度.

FCM 的目标函数为:

$$J(U,v_1,\cdots,v_c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\| x_j - v_i \|^2 \tag{2}$$

其中,$U=\{u_{ij}\}(i=1,2,\cdots,c;j=1,2,\cdots,n)$ 是隶属度矩阵,且 $0<u_{ij}<1$,$0<\sum_{j=1}^{n}u_{ij}<n$,$m>1$ 为常数,其约束为:

$$\sum_{i=1}^{c}u_{ij}=1,\ \forall j=1,\cdots,n \tag{3}$$

3　拉普拉斯加权聚类

3.1　Laplacian 加权 HCM(LHCM)

给定样本集合 $X=\{x_1,x_2,\cdots,x_n\}$,将之划分为 c 个聚类 $G_i(i=1,2,\cdots,c)$,n 为样本个数,$V=\{v_1,\cdots,v_i,\cdots,v_c\}$ 是 c 个聚类中心集合,$S=\{s_{ij}=e^{-\|x_j-v_i\|^2/t_i}\}(i=1,2\cdots,c,j=1,2,\cdots,n)$ 为第 j 个样本相对于聚类 i 的加权系数.

LHCM 目标函数为:

$$J_{LHCM}(X,S,V)=\sum_{i=1}^{c}\sum_{x_j\in G_i}s_{ij}\| x_j - v_i \|^2 \tag{4}$$

划分过的类一般用一个 $c\times n$ 的二维隶属矩阵 U 来定义. 如果第 j 个样本 x_j 属于第 i 个聚类,则 U 中的元素 u_{ij} 为 1;否则,该元素取 0. 一旦确定聚类中心 v_i,可得到使目标函数最优的 u_{ij}

$$u_{ij}=\begin{cases}1, k\neq i \text{ and } \| x_j - v_i \|^2\le\| x_j - v_k \|^2\\ 0,\text{其它}\end{cases} \tag{5}$$

通过上面定义的 u_{ij},可以看出如果样本 x_j 离聚类中心 v_i 最近,那么 x_j 属于类 i. 又由于一个给定的样本只能属于某一个聚类,所以要求此二维隶属度矩阵满足如下性质:

$$\sum_{i=1}^{c}u_{ij}=1,\ \forall j=1,\cdots,n \tag{6}$$

则由上述目标函数(4)所确定的最优类中心应满足下列等式:

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}s_{ij}x_j}{\sum_{j=1}^{n}u_{ij}s_{ij}} \tag{7}$$

LHCM 算法步骤如下:

给定数据集 $X=\{x_1,x_2,\cdots,x_n\}$,进行下列步骤,最终确定聚类中心 V:

Step 1　初始化聚类中心 $v_i(i=1,2,\cdots,c)$,最常用的方法是随机在样本中任取 c 个样本. 确定最大迭代次数 T_{max}和迭代截至误差 $\varepsilon>0$.

Step 2　用式 $S_{ij}=e^{-\|x_j-v_i\|^2/t_i}$确定权系数矩阵 S 以及式(5)确定隶属矩阵 U.

Step 3　利用式(7)调整聚类中心,如果相对上次聚类中心的改变量小于或超过最大迭代次数,则算法停止.

Step 4　返回 Step 2.

3.2　Laplacian 加权 FCM(LFCM)

LFCM 目标函数如下:

$$J_{LFCM}(U,S,v_1,v_c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}s_{ij}\| x_j - v_i \|^2 \tag{8}$$

其中 $S=\{s_{ij}=e^{-\|x_j-v_i\|^2/t_i}\}$表示权系数,$U=\{u_{ij}\}(i=$

$1,2,\cdots,c;j=1,2,\cdots,n)$ 是隶属度矩阵. 且 $0<u_{ij}<1$, $0<\sum_{j=1}^{n}u_{ij}<n$, $m>1$ 为常数,其约束为:

$$\sum_{i=1}^{c}u_{ij}=1,\ \forall j=1,\cdots,n \tag{9}$$

从而构造如下新的目标函数:

$$\overline{J}(U,S,v_1,\cdots,v_c,\lambda_1,\cdots,\lambda_n)=\sum_{i=1}^{c}\sum_{j}^{n}u_{ij}^{m}s_{ij}\|x_j-v_i\|^2+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{c}u_{ij}-1\right) \tag{10}$$

这里 $\lambda_j(j=1,2,\cdots,n)$ 是式(8)中对于 U 的 n 个约束式的拉格朗日乘子. 根据目标最优规划我们可以得到:

$$v_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}s_{ij}x_j}{\sum_{j=1}^{n}u_{ij}^{m}s_{ij}} \tag{11}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{s_{ij}\|x_j-v_i\|^2}{s_{kj}\|x_j-v_k\|^2}\right)^{1/(m-1)}} \tag{12}$$

LFCM 算法步骤如下:

给定数据集 $X=\{x_1,x_2,\cdots,x_n\}$,进行下列步骤,最终确定聚类中心 V 和隶属矩阵 U:

Step 1 初始化聚类中心 $v_i(i=1,2\cdots,c)$,即在样本集中随机选取 c 个样本. 确定最大迭代次数 T_{max}和迭代截至误差 $\varepsilon>0$.

Step 2 计算权重系数 $S=\{s_{ij}=e^{-\|x_j-v_i\|^2/t_i}\}$,从而通过式(12)初始化隶属矩阵 U.

Step 3 用式(11)计算调整聚类中心. 如果相对上次聚类中心的改变量小于阈值 ε 或超过最大迭代次数,则算法停止.

Step 4 返回 Step 2.

在算法中还可以先初始化一个任意的隶属矩阵,然后再执行迭代过程. 或者根据一些先验知识对系数矩阵 S 先进行初始化.

4 实验结果及分析

在实验中,我们在人工数据集和 UCI 标准数据集[13]上对 HCM,LHCM,FCM 和 LFCM 进行了测试,验证了本文算法的有效性. 由于考虑到随机选取的初值可能对结果产生过大影响,所以多次运行算法,然后取平均结果. FCM 和 LFCM 中统一选取参数 $m=2$,还有一些参数值会在下文具体实验中给出.

加权算法的好坏与其权系数是密切相关的,而权系数则主要是由其参数 t_i 确定,在本文中我们采用如下方法选取 t_i:

$$t_i=\begin{cases}\sigma_i^2, & x_j\in N_{ik}\\ \left(\frac{1}{c}\sum_{i=1}^{c}\sigma_i\right)^2, & \text{otherwise}\end{cases} \tag{13}$$

其中 $$\sigma_i=\frac{1}{k}\sum_{j=1}^{k}\|x_j-v_i\|^2 \tag{14}$$

式中 k 为与第 i 个聚类中心的近邻数[14], N_{ik}为第 i 个聚类中心的 k 个近邻. 由式(14)可以看出权系数能够自动适应局部结构,从而达到了间接确定权值的方法,虽然其中的 k 值选取依然需要经验性给定,但较于之前的 t_i 值选取有明显的确定性和易选择性.

4.1 人工数据集

我们使用了两个人工数据集 $D1$ 和 $D2$. 数据集 $D1$ 包括 100 个样本点,分为两类. 第一类包括 50 个样本点,以(0,0)为中心成高斯分布,第二类包括 49 个样本点,以(3,0)为中心成高斯分布,另外包括一个孤立点(200,0).

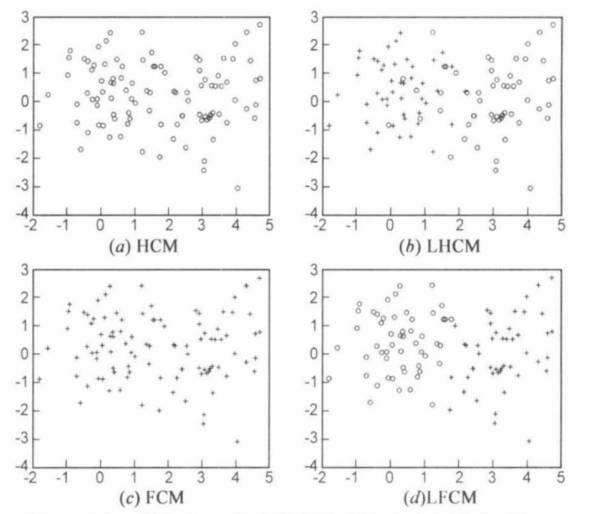

图1 对人工数据集$D1$的聚类结果(孤立点未在图中画出)

图 1 给出了 HCM、LHCM、FCM 和 LFCM 对 $D1$ 的聚类结果,图中只显示 99 个样本,孤立点未在图中显示. 可以看出,HCM、FCM 受到孤立点的干扰非常严重,把本属于两类的样本当作一类,孤立点则作为另一类. 而两种加权聚类算法则基本避免了孤立点的影响,能把两类样本近似正确分开. 可见本文提出的算法可以有效地解决孤立点的问题,由于权系数的引入,孤立点相对于聚类中心的权值很小,从而使得算法的鲁棒性得到增强,能够得到好的聚类结果.

数据集 $D2$ 是两类样本数目相差显著的情况,第一类有 25 个样本,第二类则有 125 个样本. FCM 和 LFCM 对 $D2$ 的聚类结果如图 2 所示. 从图中可以看出,FCM 把本属于第二类的部分样本分到了第一类中,而 LFCM 则正确分开了两类样本. 显然,对此数据集而言,LFCM 的聚类性能明显优越于 FCM. 通过引入权系数,使得样

本数据类内更紧,类间距离更大,从而本文提出的算法 LFCM 可以有效处理分布不均匀的样本集.

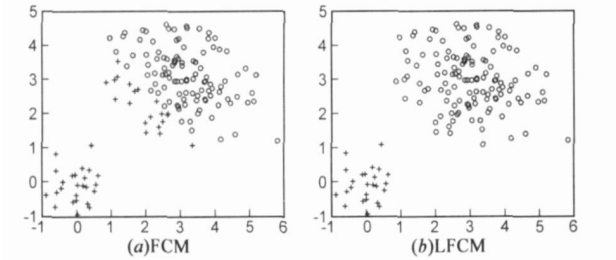

图2　对人工数据集*D*2的聚类结果

4.2　UCI 标准数据集

下面选用 UCI 标准数据集中 7 个真实的数据集进一步验证加权算法的有效性. 为了更客观的评价聚类的效果,我们采用 *F-measure*[15] 作为聚类效果的评价标准.

其定义如下:

$$F\text{-}measure = \frac{2 \times p \times r}{p + r} \tag{15}$$

其中,

$$p = \frac{t_p}{t_p + f_p},\ r = \frac{t_p}{t_p + f_n} \tag{17}$$

其中 p 为精确率, r 为反馈率, t_p 为聚类在一个类内且实际也是在一个类的样本对数目; f_p 是聚类在一个类但实际不是在一个类的样本对数目; f_n 是聚类不在一个类但实际在一个类的样本对数目.

表 1 中显示了实验所选用数据集及实验结果, *F-measure* 均为对每个数据集重复 50 次实验的平均值,加粗的数值为各种算法的最大值. 从表 1 可以看出,LFCM 的性能始终优于其它三种算法. 另外,LHCM 的性能要好于 HCM,显示了拉普拉斯加权的有效性.

表 1　HCM,LHCM,FCM 和 LFCM 算法的聚类结果

数据集	数据集样本数	类别数	属性数	HCM	LHCM	FCM	LFCM
Iris	150	3	4	0.7901	0.8344	0.8506	0.8622
Glass	214	7	9	0.4375	0.4426	0.4882	0.4928
Soybean	474	3	5	0.3855	0.4398	0.3994	0.4565
Inosphere	351	2	34	0.5811	0.5978	0.5996	0.6278
Hayes-roth	132	4	2	0.4530	0.4802	0.4958	0.4998
Balance-scale	625	3	4	0.4657	0.5813	0.5311	0.5969
Image segment	210	7	19	0.4637	0.4718	0.4687	0.4726

大多数聚类学习算法,均需人为给定聚类个数 c,且聚类结果对 c 值很敏感. 我们选择 Iris 和 Soybean 数据集来测试算法对于 c 值的敏感程度,实验结果图 3 所示. 可以看出,几乎对所有的类数 c,LHCM 和 LFCM 都取得很好的结果,而 HCM 和 FCM 则随着 c 离真实类数目远而变差. 由于权系数能够描述数据领域结构,使得在一类的数据尽量保持在一类中,不同类的仍趋于不同类,从而本文算法能够有较高的 *F-measure*.

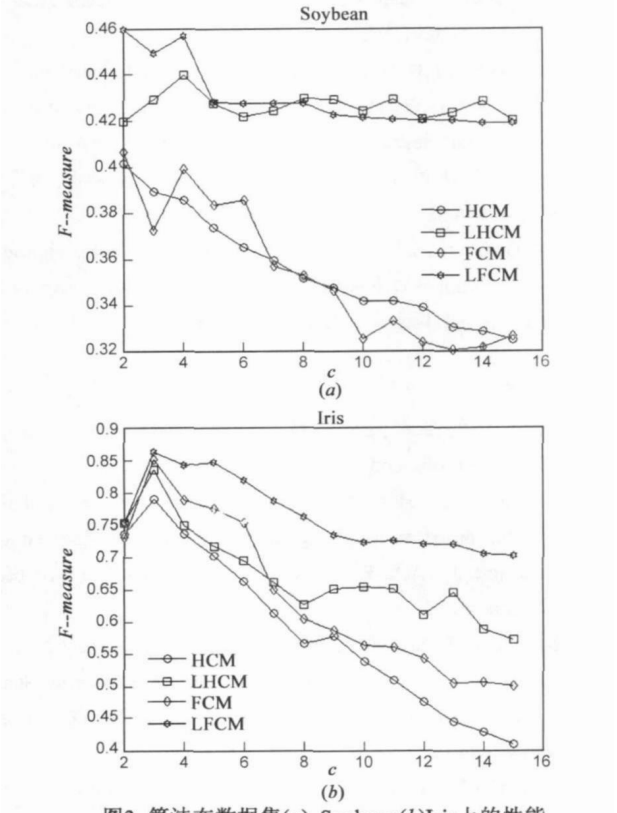

图3　算法在数据集(*a*) Soybean(*b*)Iris 上的性能

5　结束语

通过修改传统聚类算法目标函数,加入了权重系数,这种改进在实验上得到了证明,聚类结果在很大程度上得到了提高. 在本文中只讨论了全局权重系数,以后的实验中还可以考虑局部权重,提高聚类正确性,减少迭代次数. 还可以根据距离确定权重的系数范围,通过归一化的方法使得权重保持在固定范围内,从而确定不同范围权重的影响因子.

参考文献:

[1] S Ermejo, J Cabestany. The effect of finite sample size on on-line K-means[J]. Neural Computation, 2002, 48(1): 51-539.

[2] K L Wu, M S Yang. An alternative fuzzy c-means clustering algorithm[J]. Pattern Recognition, 2002, 35(10): 2267-2278.

[3] D E Gustafson, W C Kessel. Fuzzy clustering with a fuzzy covariance matrix[A]. Proc IEEE Conf Decision Control[C]. CA, 1979: 761-766.

[4] D Q Zhang, S C Chen. A comment on 'Alternative c-means clustering algorithms'[J]. Pattern Recognition, 2004, 37(2): 173-174.

[5] R N Dave. Fuzzy shell clustering and applications to circle detection in digital images [J]. Int J General Systems, 1990, 16 (4): 343-355.

[6] R O Duda, P E Hart. Pattern Classification and Scene Analysis [M]. NY: Wiley, 1973.

[7] J Z Huang, M K Ng, H Q Rong, Z C Li. Automated variable weighting in k-means type clustering [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[8] 章成志, 师庆辉, 薛德军. 基于样本加权的文本聚类算法研究 [J]. 情报学报, 2008, 27(1): 42-48.
C Z Zhang, Q H Shi, D J Xue. Document clustering algorithm based on sample weighting [J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(1): 42 - 48. (in Chinese)

[9] 高新波, 李洁, 姬红兵. 基于加权模糊 c 均值聚类与统计检验指导的多阈值图像自动分割算法 [J]. 电子学报, 2004, 32(4): 661-665.
X B Gao, J Li, H B J. A multi-threshold image segmentation algorithm based on weighting fuzzy c-means clustering and statistical test [J]. Acta Electronica Sinica, 2004, 32(4): 661-665. (in Chinese)

[10] M Belkin, P Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering [A]. Advances in Neural Information Processing Systems 14 (NIPS 2001) [C]. Cambridge, MA: MIT Press, 2002: 585-591.

[11] X F He, P Niyogi. Locality preserving projections [A]. Advances in Neural Information Processing Systems 16 (NIPS 2003) [C]. Cambridge, MA: MIT Press, 2004.

[12] D Q Zhang, Z H Zhou, S C Chen. Semi-supervised dimensionality reduction [A]. Proc 2007 SIAM Conference on Data Mining (SDM 2007) [C]. Minneapolis, MN, 2007. 629-634.

[13] C Blake, E Keogh, C J Merz. UCI repository of machine learning databases [DB/OL]. [http://www.ics.uci.edu/~mlearn/MLRepository.html], Department of Information and Computer Science, University of California, Irvine, 1998.

[14] L Zelnik-Manor, P Perona. Self-tuning spectral clustering [A]. Advances in Neural Information Processing Systems 17 (NIPS 2004) [C]. Cambridge, MA: MIT Press, 2005: 1601-1608.

[15] I H Witten, E Frank. 数据挖掘: 实用机器学习技术 [M]. 北京: 机械工业出版社, 2005. 116-118
I H Witten, E Frank. Data Mining: Practical Machine Learning Tools and Techniques [M]. Beijing: China Machine Press, 2005: 116-118. (in Chinese)

作者简介:

黄鹏飞 男, 1984 年生于江苏盐城, 硕士研究生, 主要研究领域为模式识别及应用.
E-mail: hpffph@nuaa.edu.cn

张道强 男, 1978 年生于山东枣庄, 南京航空航天大学教授, 主要从事模式识别、机器学习、数据挖掘和图像处理等领域的研究.
E-mail: dqzhang@nuaa.edu.cn