基于数据分段的K-means的优化研究

朱云贺,张春海,张 博
(中国海洋大学 信息科学与工程学院,山东 青岛 266100)

摘 要:K-means聚类算法是一种主流的迭代下降聚类算法,收敛于局部最优化状态。由于K-means随机选取k个初始聚类中心,使得聚类结果的有效性随初始输入而波动,为此文中采取一种预处理的方式来选取初始聚类中心。首先在某种范数的意义下,确定相隔最远的两个数据点之间的距离,然后采用数据分段的方法,将数据集分成k段,在每段中选取一个中心,以此来减小聚类结果随初始输入的波动。实验显示优化后的K-means有效地消除了初始输入的影响,并显著地减少了算法迭代次数和聚类误差。

关键词:聚类;K-means;PK-means;聚类中心
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2010)11-0130-03

Optimizing Research on K-means Based on Data Partition

ZHU Yun-he, ZHANG Chun-hai, ZHANG Bo
(College of Information Science & Engineering, Ocean University of China, Qingdao 266100, China)

Abstract: The K-means clustering algorithm is one kind of mainstream iterative drop clustering algorithm, which restrains in the partial optimized state. Because K-means randomly selects initial clustering center, which result in the result of clustering is obviously fluctuate along with the initial input. Thus this paper adopts the pretreatment way to select the initial clustering center. First under one kind of norm, calculate out the farthest distance, then use the method of data partition to divide the data set into k section and select a center in each section. The experiment demonstrates the optimizing K-means eliminate the initial input influence, effectively reduced the iteration number of times and clustering error.

Key words: clustering; K-means; PK-means; clustering center

0 引 言

聚类是数据挖掘的一个重要研究方向,根据在数据中发现的描述对象及其关系的信息,将数据对象分类。其目标是类内的对象之间相似,而类间的对象之间不同,类内的相似性越大,类间的相似性越小,聚类的有效性就越好[1,2]。聚类在心理学、社会学、生物学、统计学、模式识别、信息检索、机器学习等领域都扮演着重要角色[3]。为实现对数据的聚类,人们提出了很多种算法,主要有K-means算法,K-Medians算法,CLARAN算法,BITCH算法,DBCSCAN算法,CURE算法等[4~6]。

K-means聚类算法是一种主流的迭代下降聚类算法,收敛于局部最优化状态[7~9]。K-means算法在每次迭代过程中每一个聚类均用相应聚类中对象的均值来表示,而这种聚类算法是一种贪婪搜索策略,优点是它可以处理大数据集,是相对可伸缩的和高效的;缺点是聚类结果的有效性随初始输入而波动,经常会陷于局部最优。这是因为K-means首先随机选取k个初始聚类中心,将数据分割到最近的聚类中心,然后用类中数据的平均值更新类中心,重新分割数据到k个类,直到类中心不发生变化[10~12]。

为此文中描述了一种采取预处理的方式来选取初始聚类中心的PK-means算法。首先在某种范数的意义下,确定相隔最远的两个数据点之间的距离,然后采用数据分段的方法,将数据集分成k段,在每段中选取一个中心。实验显示优化后的K-means消除了初始输入的影响,有效地减少了算法迭代次数和聚类误差。

收稿日期:2010-03-16;**修回日期**:2010-06-28
基金项目:青岛市科技计划项目(08-1-3-2-jch)
作者简介:朱云贺(1986-),男,山东菏泽人,硕士研究生,研究方向为软件工程、数据库理论与应用;张春海,教授,硕士生导师,研究方向为数据库理论及应用、软件工程、工作流、构件计算。

1 优化初始聚类中心算PK-means的基本思想

定义1:簇(Cluster):一组数据对象的集合,在同一

簇中数据对象间具有相似性;而在不同簇中数据对象之间具有相异性。

定义 2:聚类分析(Cluster Analysis):给定一组数据集 $S=\{x_1,x_2,\cdots,x_n\}$,把它划分成若干聚类或簇:$\{C_1,C_2,\cdots,C_k \mid C_1 \subseteq S\}$。聚类的目标是将大量数据聚集成不同的类,使得不同类之间的数据相似性尽可能小,而各个类的内部数据相似性尽可能大。其中聚类的数目为 k,聚类中心为 $z_1,z_2,\cdots,z_k$。

定义 3:目标函数:

$$OF=\min_{z_1,z_2,\cdots,z_k}\sum_{j=1}^{n}\min_{1\leqslant p\leqslant k}\| x_j-z_p\| \tag{1}$$

由参考文献[5],当取 2 范式时,公式(1)等价于

$$OF=\sum_{i=1}^{k}\sum_{x_i\in C_i}(x_i-z_i)^2 \tag{2}$$

聚类分析需要确定 $K=\{C_1,C_2,\cdots,C_K\}$。首先需要确定 k 个聚类中心 $z_1,z_2,\cdots,z_k$,然后计算出各个数据与聚类中心的距离,最后将数据分配到最近的聚类中心里 z_i 里。因此要进行对数据的聚类分析,首先需要确定聚类中心 z_i。以下是确定聚类中心 z_i 的过程:

对公式(2)两边求导,令导数等于 0,求解 z_i 得:

$$\frac{\partial}{\partial z_i}OF=\frac{\partial}{\partial z_i}\sum_{i=1}^{k}\sum_{x_i\in C_i}(x_i-z_i)^2$$

$$=\sum_{i=1}^{k}\sum_{x_i\in C_i}\frac{\partial}{\partial z_i}(x_i-z_i)^2$$

$$=\sum_{x_i\in C_i}2(x_i-z_i)=0$$

$$\sum_{x_i\in C_i}2(x_i-z_i)=0=>$$

$$m_iz_i=\sum_{x_i\in C_i}x_i,m_i \text{ 是 } C_i \text{ 中数据的个数}$$

因此,

$$z_i=\frac{1}{m_i}\sum_{x_i\in C_i}x_i \tag{3}$$

z_i 是各个聚类的算术平均数,所以初始聚类中心也应该接近这个算术平均数。文中通过如下步骤实现这个目标。首先在某种范数的意义下,确定相隔最远的两个数据点之间的距离,然后采用数据分段的方法,将数据集分成 k 段,在第 k 段中求出算术平均数,作为第 k 个聚类中心。重复执行以上步骤,求出 k 个聚类中心。

2　优化初始聚类中心的 PK－means 算法

2.1　K－means 算法的主要步骤

1)随机选取初始聚类中心。

K－means 算法是从数据集中随机地选取 k 个中心。这种选取具有随意性和盲目性,在一定程度上决定算法的最后效果。为了改进 K－means,主要在初始聚类中心的确定上进行改进。实验验证,这种改进是有意义和必要的。

2)数据分类。

数据分类的过程实际上是根据数据的相似性对数据进行归类。首先要考虑的就是相似性的度量标准。文中采用的是在 R^n 空间中,两个数据点的相似性用某种范数意义下的大小来衡量。通过适当地选取范数可以使问题变得更加确切,相似性度量的更加准确。

3)聚类中心的调整。

根据已有的数据点的分类结果,对每一个已经有的分类,聚类中心调整为所含数据点的几何平均值。

2.2　优化的初始聚类中心算法 PK－means 描述

对初始聚类中心进行改进,改进的数学理论已经在上文中叙述。具体思想是:首先在某种范数的意义下,确定相隔最远的两个数据点之间的距离,然后采用数据分段的方法,将数据集分成 k 段,在每段中选取一个中心。

具体算法如下:

Input:数据集为 $S=\{x_1,x_2,\cdots,x_n\}$

聚类的数目为 k

Output:聚类结果 $C_1,C_2,\cdots,C_K$

聚类中心为 $z_1,z_2,\cdots,z_k$

Step1:数据预处理。将数据分段,确定初始聚类中心。

For $i=1:k$

$$M=\max_{1\leqslant i,j\leqslant n,i\neq j}\| x_i-x_j\|_2^2,d=\frac{K}{k},S_1=S,C_1=\varnothing$$

$$\| z_i\|_2^2=\max\{\| z_i\|_2^2 \mid x_j\in S_i\}$$

$$C_i=\{x_j \mid \| x_i-z_i\|_2^2\leqslant d,x_j\in S_i\}$$

$$S_{i+1}=S_i-C_i$$

$$z_i=\frac{1}{m_i}\sum_{x_i\in C_i}x_i$$

End

Step2:将数据集 $S=\{x_1,x_2,\cdots,x_n\}$ 根据中心 $z_1,z_2,\cdots,z_K$ 进行分割,得到 $C_1,C_2,\cdots,C_K$。分割标准:

$$C_i=\{x_j \mid \| x_j-z_i\|_2^2\leqslant \| x_j-z_p\|_2^2\}$$

$$p\neq i,p=1,2,\cdots,k,x_j\in S$$

Step3:调整聚类中心,得到新的聚类中心:z_1^*,$z_2^*,\cdots,z_k^*$。调整标准:$z_i^*=\frac{1}{m_i}\sum_{x_j\in C_i}x_j$

Step4:如果 $z_1\neq z_1^* \| z_2\neq z_2^* \|,\cdots,\| z_k\neq z_k^*$,转到 Step2;

Step5:输出聚类结果 $C_1,C_2,\cdots,C_K$,聚类中心为 $z_1,z_2,\cdots,z_k$。

3 仿真实验和结果分析

3.1 实验环境

实验采用 MATLAB7.0 编程环境，CPU 是 T4300，内存是 2G 的计算机上运行。实验一数据集采用 UCI(http://archive.ics.uci.edu/ml/)提供的数据集：Irish，Wine。UCI 数据库是专门提供数据挖掘的数据库，UCI 提供的数据都有明确的分类，因而可以计算算法分类的正确率。其中数据集 Iris、Wine 是国际公认的比较聚类方法效果的专门数据。数据集 Irish 由 4 维空间、150 个样本组成，总共分为 setosa、versicolor、virginica 三个聚类，每个聚类有 50 个样本数据，每个样本数据的有 4 个属性：萼片长、萼片宽、花瓣长和花瓣宽；数据集 Wine 是是由意大利食品药品分析与技术学院的 M. Forina 等人搜集整理的一个 13 维数据集，总共 178 个样本数据，分为 3 个聚类。实验二数据集是由均匀分布随机产生的规模递增的数据集，以测试 PK－means 与 K－means 在随机地、大规模数据中的性能比较。

3.2 实验结果

实验结果比较如表 1、2 和图 1 所示。

表 1 实验一结果比较表

算法	数据集					
	Iris			Wine		
	初始中心	正确率	聚类误差	初始中心	正确率	聚类误差
K－means	26,128,6	82.00%	71.73	93,39,94	55.05%	5497.32
	13,104,49	84.00%	71.42	175,96,12	65.73%	5658.14
	107,65,112	88.66%	39.47	20,167,144	58.98%	9054.70
	31.37,84	81.13%	61,71	133,23,140	60.67%	1762.14
	8,27,11	83.25%	72.63	136,59,36	56.74%	5485.01
PK－means	(6.85,3.07,5.74,2.07) (5.90,2.74,4.39,1.43) (5.00,3.41,1.46,0.24)	89.33%	39.47	1.4,14.4.1 2.5,1.48 2.32,16.8, 95.5,2.2, 2.43,.26 1.57,5.3, 1.17.6,2.8 2,1280.23	65.77%	5058.13

表 2 实验二结果比较表

随机数据规模	K－means				PK－means			
	初始误差	最终误差	运行时间	迭代次数	初始误差	最终误差	运行时间	迭代次数
10	1.935e3	6.841e2	0.005	4	1.222e3	1.323e2	0.001	2
100	2.483e5	1.497e5	0.016	16	3.3119e4	1.783e4	0.013	12
1000	6.388e6	1.300e6	0.109	29	2.183e5	1.683e5	0.078	11
10000	2.199e7	1.338e7	0.968	31	3.065e6	1.860e6	0.906	14
100000	3.036e8	1.336e8	11.891	41	4.1456e7	2.134e7	10.321	28

3.3 实验分析

表 1 显示 K－means 聚类的正确率随着不同的初始聚类中心而波动，而 PK－means 首先确定初始聚类中心，聚类的正确率稳定在较为有效的水平。表 2，图 1 显示随着数据规模的增大，PK－means 越来越有效。PK－means 算法的初始聚类误差要小于 K－means 产生的聚类误差，且迭代次数随着规模的增加明显减少。虽然增加了初始聚类中心的计算时间，但迭代次数减少，使得总体的运行时间仍然少于 K－means 的运行时间。

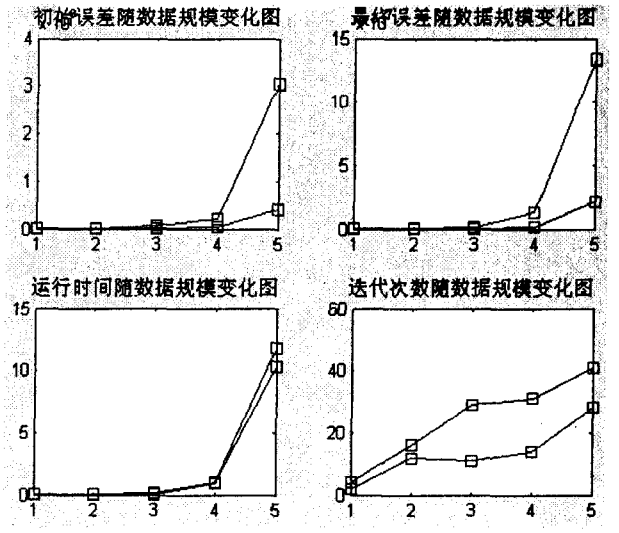

图 1 实验二结果比较图

4 结束语

提出了一种基于数据分段的 PK－means 算法，通过对初始聚类中心的优化，有效地避免了聚类结果因不同输入而产生的波动，并提高了 K－means 的聚类效率。下一步研究工作是如何提高迭代效率。

参考文献：

[1] Tan Pang－Ning, Steinbach M, Kumar V. Introduction to Data Mining[M]. 北京：人民邮电出版社，2006：324－350.

[2] Kanungo T, Mount D M, Netanyahu N S, et al. An efficient K－means clustering algorithm: analysis and implementation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881－892.

[3] Steinbach M, Karypis G, Kumar V. A comparison of document clustering techniques [C] // In: Proc of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, MA, USA: [s.n.], 2000.

[4] Su M C, Chou C H. A modified version of the K－means algorithm with a distance based on cluster symmetry[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23 (6): 674－680.

[5] Bandopadhyay S, Maulik U. An evolutionary technique based on K－Means algorithm for optimal clustering in RN[J]. Information Science, 2002, 146: 221－237.

(下转第 136 页)

从模型比较图 2 中可以很清晰看到四种机器学习算法的 Day3 模型都表现很好，SiteB 模型除了两个已经排除的算法外，另外两个都达到了很高的识别率；四个算法的 Day1 模型表现一般，都在 80% 以下，而 Day2 即使排除两个算法(NB,NBTree)也只达到平均 70%。

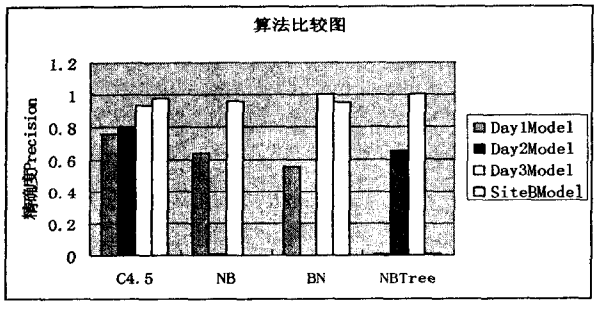

图 3 算法比较图

算法比较图 3 显示出 C4.5 算法在 P2P 流量识别方面的优越性，同时加上其较短的建模时间，所以说 C4.5 算法是很适合 P2P 的流量识别的(至少从本课题比较的这四种算法来说是如此)。其次是贝叶斯网络算法，虽然在一个数据集上表现不佳，但凭借其在其他三个数据集上的良好表现加上它拥有比 C4.5 更快的建模时间(平均提高 58%)，所以还是能肯定这个算法的。

4 结束语

提出了基于机器学习的 P2P 流量检测系统的实现方案，设计实现了基于 FCBF 方法的流量属性方案，并对网络公开数据源使用 4 种流行的机器学习算法进行了离线的分析和比较。在所比较的四种机器学习算法中，决策树算法 C4.5 在各方面的都表现较好。文中只比较了四种分类算法的性能，如前文所述，目前应用到 P2P 流量检测的机器学习算法除了分类算法外，还有无监督的学习算法，其中一个很重要的应用就是聚类分析，所以探讨聚类分析在 P2P 流量检测方面的性能是下一步研究的问题，同时利用另外的属性选择算法选取特征集再结合各种机器学习算法进行类似测试也是很有价值的研究方向。

致谢：英国剑桥大学计算机实验室的 Li Wei 为本文提供了实验数据集，在此表示由衷的感谢。

参考文献：

[1] 吴国庆.对等网络技术研究[J]. 计算机技术与发展,2008,18(7):100－104.

[2] 蒋海明,张剑英,王青青,等.P2P 流量检测与分析[J]. 计算机技术与发展, 2008,18(7):74－76.

[3] McGraw－Hill. 机器学习[M].曾华军,张银奎,等译.北京:机械工业出版社,2003.

[4] 威 滕.数据挖掘:实用机器学习技术[M]. 北京:机械工业出版社,2006.

[5] Zuev D, Moore A. Traffic classification using a statistical approach[J]. Lecture Notes in Computer Science,2005,3431:321－324.

[6] Constantinou F,Mavrommatis P. Identifying Known and Unknown Peer－to－Peer Traffic[C]// Proceedings of Fifth IEEE International Symposium on Network Computing and Applications.[s.l.]:[s.n.], 2006:93－102.

[7] Karagiannis T, Broido A, Faloutsos M, et al. Transport Layer Identification of P2P Traffic[C]//In: Proc. of ACM SIGCOMM IMC. Taormina, Sicily, Italy:[s.n.],2004:121－134.

[8] 吴 敏,王汝传. 基于主机的 P2P 流量检测与控制方案[J]. 计算机技术与发展,2009,19(10):26－30.

[9] 王 锐. P2P 流量检测技术研究[D]. 长沙:国防科学技术大学,2006.

[10] Yu Lei, Liu Huan. Feature selection for high－dimensional data: a fast correlation－based filter solution[C]// in: Proceedings of the 20th International Conference on Machine Learning (ICML'03). Washington, D.C.:[s.n.], 2003:856－863.

[11] Erman J, Mahanti A, Arlitt M, et al. Offline/realtime traffic classification using semi－supervised learning[C]//26th International Symposium on Computer Performance, Modeling, Measurements, and Evaluation.[s.l.]:[s.n.], 2007:1194－1213.

[12] Li Wei, Canini M, Moore A W, et al. Efficient application identification and the temporal and spatial stability of classification schema[J]. Computer Networks,2009,53:790－809.

(上接第 132 页)

[6] 张建辉.K－means 聚类算法研究及应用[D].武汉:武汉理工大学,2007:4－13.

[7] 张建萍,刘希玉.基于聚类分析的 K－means 算法研究及应用[J].计算机应用研究,2007(5):23－24.

[8] 吴 涛,吴 涛,尚 丽,等.一种基于聚类的交叉覆盖算法[J].计算机技术与发展,2008,18:113－116.

[9] 李士勇.蚁群优化算法及其应用研究进展[J].计算机测量与控制,2003(12):45－47.

[10] 吴 斌,傅伟鹏,郑 毅,等.一种基于群体智能的 Web 文档聚类算法[J].计算机研究与发展,2002(11):63－65.

[11] 杨黎刚,苏宏业,张 英,等.基于 SOM 聚类的数据挖掘方法及其应用研究[J].计算机工程与科学,2007(8):40－43.

[12] 袁 方,周志勇,宋 鑫.初始聚类中心优化的 k－means 算法[J].计算机工程,2007(3):76－77.