一种基于元启发式策略的迭代自学习K-Means算法

雷小锋[1] 杨 阳[1] 张 克[1] 谢昆青[2] 夏征义[3]
(中国矿业大学计算机科学与技术学院 徐州 221116)[1]
(北京大学智能科学系/视觉与听觉国家重点实验室 北京 100871)[2]
(中国人民解放军总后勤部后勤科学研究所 北京 100071)[3]

摘 要 类内误差平方和最小化的聚类准则求解是NP难问题,K-Means采用的迭代重定位方法本质上是一种局部搜索的爬山算法,因此聚类结果对初始代表点的选择非常敏感,只能保证局部最优。为此,引入元启发式策略,通过建立评估函数对K-Means初始代表点和目标函数之间的依赖关系进行近似,然后利用近似评估函数指导新的初始代表点的选择,构成一种迭代自学习框架下的K-Means算法。实验表明算法可以很好地克服K-Means对初始代表点的依赖性,获得较高质量的聚类结果。

关键词 聚类问题,K-Means算法,元启发式策略,迭代自学习框架
中图法分类号 TP181 **文献标识码** A

Metaheuristic Strategy Based K-Means with the Iterative Self-Learning Framework

LEI Xiao-feng[1] YANG Yang[1] ZHANG Ke[1] XIE Kun-qing[2] XIA Zheng-yi[3]
(School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China)[1]
(Department of Intelligence Science/National Laboratory on Machine Perception, Peking University, Beijing 100871, China)[2]
(Logistics Science and Technology Institute, P. L. A Chief Logistics Department, Beijing 100071, China)[3]

Abstract The clustering problems based on minimizing the sum of intra-cluster squared-error are known to be NP-hard. The iterative re-locating method using by K-Means is essentially a kind of local hill-climbing algorithm, which will find a locally minimal solution eventually and cause much sensitivity to initial representatives. The meta-heuristic strategy was introduced to minimize the squared-error criterion globally. Firstly, an evaluation function was built to approximate the dependency between a series of initial representatives of K-Means and the local minimal of objective criterion, and then the selection of initial representatives was done under the supervision of the evaluation function for the next K-Means. This iterative and self-learning process is called Meta-KMeans algorithm. The experimental demonstrations show that Meta-KMeans can overcome the sensitivity to initial representatives of K-Means to a great extent.

Keywords K-Means algorithm, Metaheuristic, Iterative self-learning framework

聚类分析在统计学、机器学习、数据挖掘、生物学、空间数据库、Web搜索、分布式网络、市场营销等领域得到广泛的研究和应用。形式化地,假设d维数据空间A中有n个样本,每个样本可以视为d维空间的一个点,聚类是将这n个点分组,每组就形成一个类簇,要求属于同一类簇的点尽可能相近,不同类簇间的点尽可能远离。文献[1]中对聚类方法进行了很好的综述,将主要的聚类算法概括为划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法。

KMeans是最典型的一种划分方法,本质上可视为基于局部搜索的爬山算法,因此K-Means对初始代表点的选择非常敏感,且只能保证收敛到局部最优解。本文引入元启发式策略和迭代自学习框架,提出Meta-KMeans聚类算法,很大程度上克服了K-Means对初始代表点的依赖性,获得高质量的聚类结果。

假设用x表示一组初始代表点,而所有可能的x组成的搜索空间用X表示,$\|X\|=C_n^k$,则对于任意初始代表点x,K-Means算法可以视为一个映射f:

$\forall x\in X, f(x)\rightarrow MinJ_e$

这里,K-Means算法将初始代表点x映射为一个实数$MinJ_e$,其中J_e是误差平方和,并将初始代表点改进为一组类簇中心x_c。实际上,从初始代表点x到类簇中心x_c之间存在一条搜索路径,以该路径上任意位置作为初始代表点,均会得到最终映射值$MinJ_e$。K-Means算法对初始代表点选择的敏感性说明在代表点x和$MinJ_e$之间存在很强的依赖性,能否把这种依赖性建模为一个评估函数,并将已经获得的一组$(x, MinJ_e)$对作为训练集,对该评估函数进行学习;最终利用评估函数来指导初始代表点的选择。这实际上是一种元启发式的思想。此外,评估函数可以指导选择初始点,而由该初始点算法

到稿日期:2008-08-29 返修日期:2009-03-09 本文受中国矿业大学科技基金(OD080313),国家863高技术研究发展计划(2006AA12Z217)资助。

雷小锋(1975—),男,博士,副教授,主要研究方向为数据库、数据挖掘、机器学习图像处理等,E-mail:leiyunhui@gmail.com。

又会生成新的训练样例来改进评估函数，是一种迭代自学习的框架。元启发式策略和迭代自学习框架构成 Meta-KMeans 算法的核心。

本文第 1 节对 Meta-KMeans 算法的元启发式策略和迭代自学习框架进行说明，并给出具体的算法描述；第 2 节分别从聚类结果的有效性和效率两方面进行实验比较；最后是结论和展望。

1 Meta-KMeans 算法

1.1 基本思路

自从 1983 年模拟退火(SA，Simulated Annealing)算法[2]提出以来，元启发式方法的研究迅速成为最主要的研究热点，先后提出了禁忌搜索(Tabu Search)[3,4]、蚁群系统(Ant Systems)[5]、STAGE[6]等新方法，此外传统的遗传算法也因此焕发青春。这些方法具有一个共同的框架：从随机的可行初始解出发，利用迭代改进的策略来逼近问题的最优解。从这个意义下看 K-Means 算法也遵循元启发式方法的基本框架，但是与局部搜索的爬山法类似，算法很容易陷入局部最优解，如果没有中六合彩的极好运气，通常这个局部最优解相当的平凡。为此，我们借鉴 STAGE 算法的启发式思想，建立了评估函数来对可行解与目标函数之间的依赖关系进行近似。

已知 K-Means 算法的优化目标是最小化误差平方和的准则函数(记作 MSE)，具体定义如下：

$$MSE=\mathrm{Min}\sum_{j=1}^{k}\sum_{x\in C_j}\|x-Mean(C_j)\|$$

定义 1(评估函数 f_E) 对任意一组初始代表点 $sx=\{x_1,\cdots,x_k\}$，可以定义评估函数 $f_E(sx)$，表示从状态 sx 开始，K-Means 算法可期望的最小准则函数值 MSE。

很明显，若对函数 f_E 的行为和性质完全了解，则聚类就不存在问题，但是在初始时对该评估函数一无所知。这里借鉴 STAGE 算法的思路，可以通过多次执行 K-Means 算法得到一些非全局最优解，并以这些非最优解作为训练样例，学习出一个评估函数的近似版本(记作 $\hat{f}_E$)，然后可以利用该近似函数指导新的初始代表点的选择。这实际上是一种迭代自学习的框架，如图 1 所示。

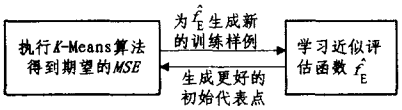


图 1 Meta-KMeans 算法的迭代自学习框架

定义 2(训练样例) 给定的一组初始代表点 sx_0，K-Means 算法会得到期望的最小准则函数值 MSE，则二元组 (sx_0,MSE)就是近似评估函数$\hat{f}_E$ 的一组训练样例。

定义 3(迭代轨迹) K-Means 算法在每次迭代过程中持续改进当前的代表点，最终得到的一组代表点就是局部最优的类簇代表点，与之相应的是期望的最小准则函数值 MSE。对于给定的一组初始代表点 sx_0，K-Means 的整个迭代过程实际上对应于一个由多组代表点构成的轨迹，称为 K-Means 算法的迭代轨迹，记作$\langle sx_0,sx_1,\cdots,sx_n\rangle$，表示该迭代轨迹从 sx_0 开始，到 sx_n 结束。

显然，对于迭代轨迹 $T=\langle sx_0,sx_1,\cdots,sx_n\rangle$，可得训练样例$(sx_0,MSE)$和$(sx_n,MSE)$。此外，根据 K-Means 算法的性质易得结论：以轨迹序列 T 中任意一组代表点 sx_i 作为初始代表点 K-Means 均可到达 sx_n，并最终得到期望的 MSE。可以看出，从迭代轨迹 T 可以生成 n 个训练样本：(sx_i,MSE)，$i=0,\cdots,n$。

直接以一组初始代表点作为训练样例学习到的近似评估函数$\hat{f}_E$ 有 $d\times k$ 个自变量。为了简化问题，这里对初始代表点所在的样本空间进行离散编码。

定义 4(样例编码) 将样本空间的每一维划分为 2^t 个区间，则整个数据空间被划分为 2^{dt} 个网格，每个网格可以被编码为长度为 dt 的二进制位串。对于任意作为训练样例的代表点，均可计算出其所在网格，并将该网格编码的十进制数值称为是该训练样例的编码

$$t=(\log_2(\sqrt{n}+1))+1,$$

通过对训练样例进行编码，将近似评估函数$\hat{f}_E$ 的自变量数目降为 k 个。

1.2 算法描述

根据图 1 给出的 Meta-KMeans 算法的迭代自学习框架，具体的算法描述如下：

算法 1 Meta-KMeans 算法。

```
Function Meta_K Means(SetOfPoints,k,Limits)
        Return k个类簇的中心
输入：待聚类的样本集 SetOfPoints;
      最终的类簇数目 k;
      迭代次数的限制 Limits;
局部变量：
      current，k 个初始代表点组成的点集;
      trajectory，表示 K-Means 迭代的轨迹序列的点集;
      collection，由(k 个初始代表点的离散编码，MSE)二元组构成
      的训练样例集;
//生成初始代表点
current = Make_InitialPoints(SetOfPoints,k);
Do While(迭代次数 <= Limits)
    (1) //以 current 为初始代表点，执行 K-Means
        //返回 MSE 和中心点的迭代轨迹
    [MSE,trajectory] = K-Means(SetOfPoint,current);
    (2) //对轨迹上的每组中心点进行离散编码，
        //并将每组中心点编码与相应的 MSE
        //组成的二元组插入集合
    collection. Insert(trajectory,MSE);
    (3) //利用训练样例对近似评估函数进行拟合
    Fittor = Fitting(collection,type);
    (4) //根据近似评估函数生成新的初始代表点
    current = StochasticHillClimbing(Fittor);
    (5) If(多次迭代的 MSE 没有改进) Break;
Loop
Return current;
End Function
```

当迭代次数超过给定阈值，或者多次迭代的 MSE 没有得到改进时 Meta-KMeans 算法停止。

2 实验结果和性能比较

Meta-KMeans 算法在 K-Means 算法的基础上扩展实现，因此其时空复杂度是 K-Means 算法的常数倍。本节主要是通过实验比较来研究 Meta-KMeans 算法的有效性。实验数据包括两组，一组通过模拟生成，另一组采用来自 GLCF (Global Land Cover Facility)的覆盖美国内华达州南部的 ETM 遥感影像(影像编号 N-11-35)。实验的硬件环境为 P4 2.0GHz 处理器和 1.0G 内存，软件平台为 MS Windows XP

professional 操作系统，所有代码均利用 Matlab 7.04 和 Visual C++6.0 实现。

模拟数据有 5 组(Data01～Data05)，分别由 3×3，4×4，5×5，6×6，7×7 个类簇组成，每个类簇是包含 100 个高斯分布采样点，具体如图 2 所示。由于篇幅所限，仅仅示出 Data03 的聚类结果，如图 3—图 5 所示：图 3 为 K-Means 算法的聚类结果，图 4 为采用线性拟合的 Meta-KMeans 的聚类结果，图 5 为采用二次拟合的 Meta-KMeans 的聚类结果。其余各组模拟数据的聚类结果质量(以 *MSE* 来衡量)和相应的迭代收敛次数在表 1 中给出。

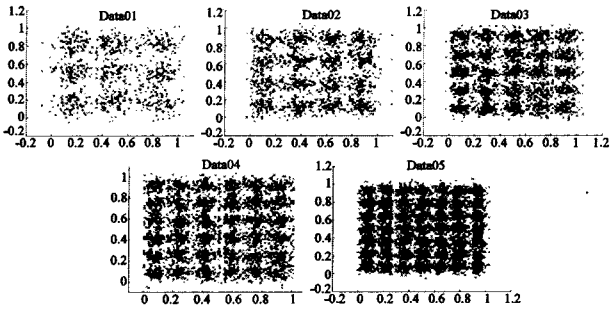


图 2　几组评估聚类算法的模拟数据

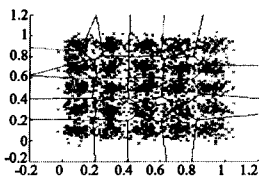


图 3　K-Means 算法的聚类结果

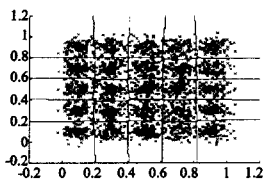


图 4　采用线性拟合的 Meta-KMeans 的聚类结果

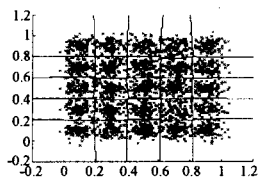


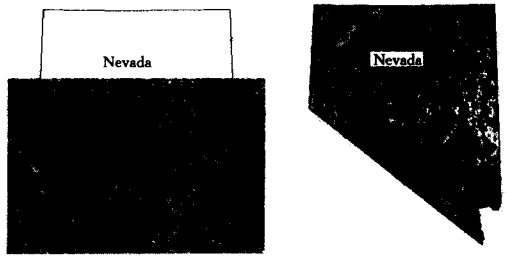
图 5　采用二次拟合的 Meta-KMeans 的聚类结果

表 1 中，MSE of K-Means 是多次 K-Means 聚类后得到的最好结果，重复次数与二次算法迭代次数相同。可以看出，从 Data02—Data05 4 个比较复杂的数据集开始，K-Means 和 Meta-KMeans 算法的结果差异逐渐显现出来，Meta-KMeans 算法明显地对聚类结果的质量有所改善，趋于最优解。实验结果表明，无论采用线性函数还是二次函数对近似评估函数进行拟合，其聚类结果差异不大，但线性拟合更为简单、效率更高。

表 1　几组模拟数据的 K-Means 和 Meta-KMeans 聚类结果比较

数据集	MSE of *K*-Means	MSE of Meta-*KMeans* & 线性函数	线性算法的迭代次数	MSE of Meta-*KMeans* & 二次函数	二次算法的迭代次数
Data01	0.01227	0.01227	14	0.01227	19
Data02	0.00823	0.00704	17	0.00704	33
Data03	0.00540	0.00440	26	0.00440	51
Data04	0.00385	0.00326	37	0.00326	73
Data05	0.00266	0.00228	50	0.00226	97

从内华达州遥感影像数据中抽样获得 1260 个样本点，每个点具有 RGB 三色属性值。此外，为了评估聚类结果的精度，实验采用了来自 MRLC(Multi-Resolution Land Characterization Project)的美国内华达州土地覆盖专题数据集作为验证标准，该数据集在互联网上公开：http://keck.library.unr.edu/data/nvlandcover/nvlc.html。图 6 分别是美国内华达州的 ETM 影像和土地覆盖专题数据。



(a) 内华达州南部遥感影像图　　(b)内华达州土地利用图

图 6　内华达州 ETM 影像和土地利用数据

由于样本点的类别归属已知，故可以采用文献[7]使用的 *F* 统计指标以及 *MSE* 指标来评价聚类结果的质量，*F* 指标值越大说明聚类结果质量越好。*F* 指标的定义基于召回率 *R* 和准确率 *P*，其定义如下：

$$F=\frac{2PR}{P+R},$$

$$R=\frac{TP}{TP+FN}=\frac{\sum_{i=1}^{k}TP_i}{\sum_{i=1}^{k}(TP_i+FN_i)},$$

$$P=\frac{TP}{TP+FP}=\frac{\sum_{i=1}^{k}TP_i}{\sum_{i=1}^{k}(TP_i+FP_i)}$$

其中，k 为类别数，TP_i 表示算法认定属于第 i 类且确实属于第 i 类的样本数，FP_i 表示算法认定属于第 i 类而实际上不属于第 i 类的样本数，FN_i 表示算法认定不属于第 i 类而实际上属于第 i 类的样本数，事实上还有 TN_i 表示算法认定不属于第 i 类且确实不属于第 i 类的样本数。最终的实验结果如表 2 所列。可以看出，采用线性拟合的 Meta-KMeans 算法无论从 *MSE* 还是 *F* 指标上均优于 K-Means 算法。

表 2　内华达 ETM 数据的 K-Means 和 Meta-KMeans 聚类结果比较

K-Means 算法					
MSE=1491.760					
F=0.442					
	C_1	C_2	C_3	C_4	C_5
C_1	134	42	2	0	8
C_2	90	93	5	0	23
C_3	22	42	22	0	127
C_4	11	3	2	28	302
C_5	6	11	7	0	280
Meta-KMeans & 线性拟合					
MSE = 1383.614					
F = 0.510					
	C_1	C_2	C_3	C_4	C_5
C_1	28	232	2	3	11
C_2	0	362	7	11	6
C_3	0	115	21	42	21
C_4	0	23	6	93	87
C_5	0	8	2	42	138

结束语　K-Means 是一种典型的基于划分的聚类方法。从搜索的角度看，K-Means 可视为基于局部搜索的爬山算法，因此对初始代表点(爬山起始点)的选择非常敏感，只能保证收敛到局部最优解。借鉴 STAGE 算法的思想，本文引入元启发式策略和迭代自学习框架，建立评估函数对可行解与目标函数之间的依赖关系进行近似，从而提出 Meta-KMeans 聚

类算法，很大程度上克服了 K-Means 对初始代表点的依赖性，可以获得较高质量的聚类结果。但是，无疑 Meta-KMeans 算法的时间复杂度较 K-Means 算法要高很多，本文下一步工作将就该问题展开研究。

参 考 文 献

[1] Han J W, Kamber M. Data Mining: Concepts and Techniques. 2nd ed[M]. Morgan Kaufmann Publishers, 2001: 223-250

[2] Kirkpatrick S, Gelatt C D Jr, Vecchi M P. Optimization by Simulated Annealing[J]. Science, 1983, 220(4598): 671-680

[3] Glover F. Tabu search-Part I[J]. ORSA Journal on Computing, 1989, 1(3): 190-206

[4] Glover F. Tabu search-Part II[J]. ORSA Journal on Computing, 1990, 2(1): 4-32

[5] Dorigo M, Blum C. Ant colony optimization theory: a survey[J]. Theoretical Computer Science, 2005, 344(2/3): 243-278

[6] Boyan J A, Moore A W. Learning Evaluation Functions for Global Optimization and Boolean Satisfiability[C]// Jack Mostow, Chuck Rich, eds. Proc. of the 15th National Conference on Artificial Intelligence. CA, USA: AAAI Press, 1998

[7] Sebastiani F. A tutorial on automatic text categorization [C] // Anala Amandi, Ricardo Z, eds. Proc. of the 1st Argentinean Symposium on Artificial Intelligence. Buenos Aires, 1999: 7-35

(上接第 174 页)

动的低层特征分析和序列、对象探测；进一步探索分析多通道低层特征以抽取更准确、有效的语义内容表示；同时，本体的自动构建也是未来研究的一个重要方向。

参 考 文 献

[1] Chang S-F. The holy grail of content-based media analysis[J]. IEEE Multimedia, 2002, 9(2): 6-10

[2] Yoshitaka A, Ichikawa T. A survey on content-based retrieval for multimedia databases[J]. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(1): 81-93

[3] Hanjalic A, Xu L Q. Affective video content representation and modeling[J]. IEEE Transactions on Multimedia, 2005, 7(1): 143-154

[4] Muller-Schneiders S, Jager T, Loos H S, et a. Performance evaluation of a real time video surveillance system[C]// 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Oct. 2005: 137-143

[5] Hua X S, Lu L, Zhang H J. Automatic music video generation based on the temporal pattern analysis[C]// 12th Annual ACM International Conference on Multimedia. October 2004

[6] Informedia-II: Auto-Summarization and Visualization over Multiple Video Documents and Libraries[R]. September 2001. http://www.informedia.cs.cmu.edu

[7] Resource description framework. Technical report [EB/OL]. W3C. http://www.w3.org/RDF/, Feb. 2004

[8] Web ontology language (OWL) [EB/OL]. Technical report. W3C. http://www.w3.org/2004/OWL/, 2004

[9] Leonardi R, Migliorati P. Semantic index of multimedia documents[J]. IEEE Multimedia, 2002, 9(2): 44-51

[10] Ekin A, Tekalp A M, Mehrotra R. Automatic soccer video analysis and summarization[J]. IEEE Transactions on Image Processing, 2003, 12(7): 796-807

[11] Yu X, Xu C, Leung H, et al. Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video[C]// ACM Multimedia 2003. Berkeley, CA (USA), Nov. 2003, 3: 11-20

[12] Xu H X, Chua T-S. Fusion of AV features and external information sources for event detection in team sports video[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2006, 2(1): 44-67

[13] Reidsma D, Kuper J, Declerck T, et al. Cross document ontology based information extraction for multimedia retrieval[C]// Supplementary Proceedings of the ICCS03. Dresden, July 2003

[14] Mezaris V, Kompatsiaris I, Boulgouris N, et al. Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(5): 606-621

[15] Jaimes A, Tseng B, Smith J. Modal keywords, ontologies, and reasoning for video understanding[C]// International Conference on Image and Video Retrieval (CIVR 2003). July 2003

[16] Jaimes A, Smith J. Semi-automatic, data-driven construction of multimedia ontologies[C]// Proc. of IEEE Int'l Conference on Multimedia & Expo. 2003

[17] Bertini M, DelBimbo A, Torniai C. Enhanced ontoloies for video annotation and retrieval[C]// ACM MIR'2005. Singapore, November 2005

[18] Sadlier D A, O'Connor N E. Event Detection in Field Sports Video Using Audio-visual Features and A SVM [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2005, 15(10): 1225-1233

[19] Dasiopoulou S, Papastathis V K, Mezaris V, et al. An Ontology Framework for Knowledge-Assisted Semantic Video Analysis and Annotation [C]// Proc. 4th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2004) at the 3rd International Semantic Web Conference (ISWC 2004). November 2004

[20] Strintzis J, Bloehdorn S, Handschuh S, et al. Knowledge representation for semantic multimedia content analysis and reasoning[C]// European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology. Nov. 2004

[21] Kompatsiaris I, Mezaris V, Strintzis M G. Multimedia content indexing and retrieval using an object ontology[A]// Stamou G, ed. Multimedia Content and Semantic Web Methods, Standards and Tools. New York: Wiley, 2004

[22] Artale A, Franconi E. A temporal description logic for reasoning about actions and plans[J]. Journal of Artificial Intelligence Research, 1998, 9: 463-506

[23] Chen J Y, Li Y H, Lao S Y, et al, Detection of Scoring Event in Soccer Video for Highlight Generation[R]. National University of Defense Technology, 2004

[24] Pan Hao, van Beek P, Sezan M I. Detection of Slowmotion Replay Segments in Sports Video for Highlights Generation[C]// Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'01). Salt Lake City, UT, USA, May 2001

[25] Bai Liang, Hu Yanli, Lao Songyang, et al. Feature Analysis and Extraction for Audio Automatic Classification[C]// IEEE SMC 2005. Hawaii USA, October 2005

[26] Zhou W, Dao S, Jay Kuo C-C. On-line knowledge and rule-based video classification system for video indexing and dissemination [J]. Information Systems, 2002, 27(8): 559-586

[27] MPEG-7 Overview[OL]. http://www.chiariglione.org, October 2004