南京航空航天大学

论文集

（二〇〇七年） 第18册

信息科学与技术学院

（第2分册）

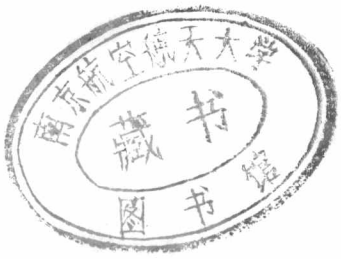
南京航空航天大学科技部编

二〇〇八年三月

(18)

信息科学与技术学院

043 系



信息科学与技术学院2007年发表论文目录-3.1

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
1	胡彩平 秦小麟	博士生 教 授	043 043	一种改进的基于密度的抽样聚类算法	中国图象图形学报	2007.12.11	
2	胡彩平 秦小麟	博士生 教 授	043 043	空间数据挖掘研究综述	计算机科学	2007.34.05	
3	胡彩平 秦小麟	博士生 教 授	043 043	一种改进的空间聚类算法	模式识别与人工智能	2007.20.03	
4	胡彩平 秦小麟	博士生 教 授	043 043	Spatial Prediction Models for Mining Spatial Data	ICIT2007年Proceedings会议交流		
5	胡彩平 秦小麟	博士生 教 授	043 043	A Novel Spatial Clustering Algorithm with Sampling	2007年Lecture Notes in Artificial Intelligencn会议交流		
6	管致锦 秦小麟	博士生 教 授	043 043	Network Structure Cascade for Reversible Logic	2007年International conference nature computation会议交流		
7	管致锦 秦小麟	博士生 教 授	043 043	Reversible Gate Cascade Based on SOP of Optimization	2007年Proceedings of the International Conference on Complex Systems and Applications会议交流		
8	郑吉平 秦小麟 孙 瑾	博士生 教 授 博士生	043 043 043	Data dependency based recovery approaches in survival database systems	2007年International conference on computation Science会议交流		
9	崔新春 秦小麟 郑吉平 管致锦	博士生 教 授 博士生 博士生	043 043 043 043	A Chaotic Sequence Based Aigorithm for Watermarking Relational Databases	2007年 The Second International Conference on Complex Systems and Applications-Modeling,Control and Simulations会议交流		
10	许建秋 秦小麟	硕士生 教 授	043 043	一种基于NWN的交通网距离算法	计算机科学(增刊)	2007.34.10B	
11	朱朝晖 肖文洁	教 授 博士生	043 043	Two Representation Theorems for Non-monotonic Inference Relations	Journal of Logic and Computation	2007.17.04	
12	朱朝晖 张 蓉	教 授 硕士生	043 043	An algebraic Characterization of equivalent preferential models	Journal of symbolic logic	2007.72.03	
13	张晋津 朱朝晖	博士生 教 授	043 043	A modal characterization of Lambda-bisimilarity	International journal of software and Informatics	2007.01.01	
14	王钦玉 马维华	硕士生 教 授	043 043	一种低成本嵌入式互联系统的设计与实现	计算机应用	2007.27.06	
15	王钦玉 马维华	硕士生 教 授	043 043	基于LPC2292的嵌入式Etherned-CAN转换器	南京航空航天大学学报	2007.39.05	
16	郑中业 马维华	硕士生 教 授	043 043	一种带优先权可扩展的数字网络视频监控系统	装备制造技术	2007.00.03	
17	郑中业 马维华	硕士生 教 授	043 043	通用嵌入式Web服务器模块的设计与实现	石家庄铁道学院学报	2007.20.02	
18	王 辉 马维华	硕士生 教 授	043 043	蓝牙高级音频分发框架协议的研究与实现	装备制造技术	2007.00.04	
19	白延敏 马维华	硕士生 教 授	043 043	基于IBM服务器的硬RAID和纯软双机热备系统	装备制造技术	2007.00.04	
20	解书钢 马维华	硕士生 教 授	043 043	基于CAN总线井下人员定位系统的设计	装备制造技术	2007.00.10	

信息科学与技术学院2007年发表论文目录-3.2

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
21	陈松灿 王　喆	教 授 博士生	043 043	Matrix-patten-oriented Ho-Kashyap classifier with regularization learning	Pattern Recognition	2007.40.05	
22	王　喆 陈松灿	博士生 教 授	043 043	New Least Squares Support Vector Machines Based on Matrix Patterns	Neural Processing Letters	2007.26.01	
23	杨绪兵 陈松灿	博士生 教 授	043 043	局部化的广义特征值最接近支持向量机	计算机学报	2007.30.08	
24	杨绪兵 陈松灿 潘志松	博士生 教 授 副教授	043 043 外	Multicategory Nonparallel Proximal Support Vector Machine	2007年International conference on Intelligent Systems and Knowledge Engineering会议交流		
25	孙廷锴 陈松灿	博士生 教 授	043 043	Locality preserving CCA with applications to data visualization and pose estimation	Image and Vision Computing	2007.25.05	
26	孙廷锴 陈松灿	博士生 教 授	043 043	Class label versus sample label-based CCA	Applied Mathematics and Computation	2007.185.01	
27	蔡维玲 陈松灿 张道强	博士生 教 授 副教授	043 043 043	Robust fuzzy relational classifier incorporating the soft class labels	Pattern Recognition Letters	2007.28.16	
28	蔡维玲 陈松灿 张道强	博士生 教 授 副教授	043 043 043	Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation	Pattern Recognition	2007.40.03	
29	蔡维玲 陈松灿 张道强	博士生 教 授 副教授	043 043 043	Enhanced fuzzy relational classifier with relational training samples	2007年International Conference on Wavelet Analysis and Pattern Recognition会议交流		
30	薛　晖 陈松灿	博士生 教 授	043 043	Alternative robust local embedding	2007年International Conference on Wavelet Analysis and Pattern Recognition会议交流		
31	陈　斌 冯爱民 陈松灿	博士生 讲 师 教 授	043 043 043	基于单簇聚类的数据描述	计算机学报	2007.30.08	
32	王　颖 陈松灿 张道强 杨绪兵	硕士生 教 授 副教授 博士生	043 043 043 043	模糊K-平面聚类算法	模式识别与人工智能	2007.20.05	
33	刘　俊 陈松灿 谭晓阳 张道强	讲 师 教 授 副教授 副教授	043 043 043 043	Efficient Pseudo-inverse Linear Discriminant Analysis and its Nonlinear Form for Face Recognition	International Journal of Pattern Recognition and Artificial Intelligence	2007.21.08	
34	刘　俊 陈松灿 谭晓阳 张道强	讲 师 教 授 副教授 副教授	043 043 043 043	Comments on “Efficient and Robust Feature Extraction by Maximum Margin Criterion”	IEEE Transactions on Neural Networks	2007.18.06	

信息科学与技术学院2007年发表论文目录-3.3

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
35	刘 俊 陈松灿 周志华 谭晓阳	讲 师 教 授 教 授 副教授	043 043 外 043	Single Image Subspace for Face Recognition	2007年IEEE International Workshop on Analysis and Modeling of Faces and Gestures (Inconjunction with ICCV 2007)会议交流		
36	吴再达 毛宇光	硕士生 副教授	043 043	一种分布式数据流动态负载平衡策略的研究	计算机科学	2007.34.10B	
37	刘宁钟	副教授	043	三维条码的编码理论和设计	计算机学报	2007.30.04	
38	曹 雁 刘宁钟	硕士生 副教授	043 043	Fast algorithm for intra-prediction mode selection in H.264/AVC based on Pan algorithm	2008年第五届SPIE多谱图像处理与模式识国际学术会议交流		
39	崔子冠 刘宁钟	硕士生 副教授	043 043	A Frame-layer rate control aigorithm for H.264 based on improved frame MAD	2008年第五届SPIE多谱图像处理与模式识国际学术会议交流		
40	杨 阳 万麟瑞	硕士生 副教授	043 043	改进的A*算法在现代电子商务运输中的运用	科技信息	2007.00.04	
41	杨 阳 万麟瑞	硕士生 副教授	043 043	面向Petri网的电子商务车辆调度模型设计	职业圈	2007.00.02	
42	孙 敏 万麟瑞	硕士生 副教授	043 043	面向Agent的模糊评标软件模型研究	计算机工程与设计	2007.28.15	
43	王 珩 万麟瑞	硕士生 副教授	043 043	工作流元过程定义软件构架研究	计算机应用研究(增刊)	2007.24(下)	
44	殷 翔 万麟瑞	硕士生 副教授	043 043	AOP建模方法及语义表示机制研究	计算机应用研究(增刊)	2007.24.	
45	曹 中 章 勇	硕士生 副教授	043 043	二进离散小波神经网络在传感器逆向建模中的应用	电脑知识与技术	2007.03.03	
46	曾丽敏 章 勇	硕士生 副教授	043 043	基于自适应用户模型的RSS内容过滤	江南大学学报	2007.06.06	
47	曾丽敏 章 勇 Robin Qiu	硕士生 副教授 教 授	043 043 外	Adaptivc Uscr Profiling in Enhancing RSS-bascd Information Scrviccs	2007年IEEE/INFORMS Int.Conf.on Service Operations and Logistics, and Informatics,Philadelphia会议交流		
48	曾丽敏 章 勇 邱广华	硕士生 副教授 教 授	043 043 外	Towards Assistive Information Services	2007年IEEE International Conference on Grey Systems and Intelligent Services会议交流		
49	瞿新南 徐 涛	硕士生 教 授	043 外	基于Agent的动态软件体系结构研究	计算机应用研究	2007.24.	
50	徐春燕 徐 涛	硕士生 教 授	043 外	面向实时数据仓库的ETL研究	计算机应用研究(增刊)	2007.24.	
51	陈建辉 徐 涛	硕士生 教 授	043 外	抵御Dos攻击的认证协议安全方案	计算机工程与设计	2007.28.22	
52	顾 梅 徐 涛	硕士生 教 授	043 外	基于ASM的彩色视频中运动物体的定位	计算机应用	2007.27.01	
53	裔传俊 徐 涛	硕士生 教 授	043 外	基于标准差的快速分形图像编码	小型微型计算机系统	2007.28.12	

信息科学与技术学院2007年发表论文目录-3.4

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
54	邹　庆 徐　涛	硕士生 教　授	043 外	基于实时数据仓库的OLAP研究	计算机与现代化	2007.11.00	
55	李领治 郑洪源 丁秋林	博士生 副教授 教　授	043 043 043	一种基于改进蚁群算法的选播路由算法	电子与信息学报	2007.29.02	
56	陈　明 周　良	硕士生 副教授	043 043	移动Agent在炮兵侦察目标信息处理中的应用研究	中国制造业信息化	2007.36.21	
57	王金栋 周　良 张　磊 丁秋林	博士生 副教授 博士生 教　授	043 043 043 043	基于分枝路径的连续查询降载算法	应用科学学报	2007.25.01	
58	徐新卫 周　良 丁秋林	博士生 副教授 教　授	043 043 043	Web服务失效处理的反射中间件技术应用与实现	系统工程与电子技术	2007.29.08	
59	徐新卫 周　良 丁秋林	博士生 副教授 教　授	043 043 043	Web主动服务中基于混合挖掘的用户意图辨识	吉林大学学报(工学版)	2007.37.02	
60	徐新卫 王有远 曹永忠 丁秋林	博士生 博士生 博士生 教　授	043 043 043 043	基于反射技术的Web服务失效处理	计算机工程与应用	2007.43.12	
61	徐新卫 丁秋林	博士生 教　授	043 043	基于QoS上下文的Web服务动态合成*	华南理工大学学报(自然科学版)	2007.35.01	
62	徐新卫 姜　楠 杨淑群 丁秋林	博士生 博士生 博士生 教　授	043 043 043 043	一种基于小世界网络的Web服务智能路由	计算机科学	2007.34.09	
63	徐新卫 曹永忠 谢　强 丁秋林	博士生 博士生 讲　师 教　授	043 043 043 043	基于分层混合专家神经网路的Web服务失效检测机制	信息与控制	2007.36.06	
64	朱保平 周　良	博士生 副教授	043 043	基于细胞自动机的公钥密码体制研究	南京理工大学学报(自然科学版)	2007.31.05	
65	谢　强 于雅丽 丁秋林	讲　师 硕士生 教　授	043 043 043	基于对象池和数据缓存技术的Web-OLAP系统	应用科学学报	2007.25.02	
66	谢　强 冯桂焕 孙正兴	讲　师 博士生 教　授	043 043 外	基于上下文的在线草图识别方法	计算机科学	2007.34.03	
67	谢　强 张　磊 周　良	讲　师 博士生 副教授	043 043 043	基于改进粒子群优化算法的Ontology划分方法	华南理工大学学报(自然科学版)	2007.35.09	

信息科学与技术学院2007年发表论文目录-3.5

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
68	冯桂焕 孙正兴 谢　强	博士生 教　授 讲　师	043 外 043	基于手绘草图的方案设计CAD系统	吉林大学学报(信息科学版)	2007.19.06	
69	周日贵 谢　强 丁秋林	博士生 讲　师 教　授	043 043 043	多模式高概率量子搜索算法	南京航空航天大学学报	2007.39.02	
70	周日贵 姜　楠 丁秋林	博士生 博士生 教　授	043 043 043	Multi-Pattern Recongnition Based on Grover's Algorithm*	Chinese Journal of Electronics	2007.16.04	
71	周日贵 丁秋林	博士生 教　授	043 043	Quantum M-P Neural Network	International Journal of Theoretic Physics	2007.46.00	
72	周日贵 曹永忠 杨淑群 徐新卫	博士生 博士生 博士生 博士生	043 043 043 043	Quantum Storage Network	2007年Third Intemational Conferenec on Natural Computation(ICNC2007)会议交流		
73	周日贵 姜　楠 杨淑群 丁秋林	博士生 博士生 博士生 教　授	043 043 043 043	Feature Selection based on QFT	2007年Third Intemational Conferenec on Natural Computation(ICNC2007)会议交流		
74	於在稳 谢　强 丁秋林	硕士生 讲　师 教　授	043 043 043	基于MDA的软件逆向工程研究及应用	中国制造业信息化	2007.36.21	
75	黄卫东 谢　强 丁秋林	博士生 讲　师 教　授	043 043 043	产品设计中基于本体理论的知识管理研究	信息与控制	2007.36.05	
76	黄卫东 王有远 谢　强 丁秋林	博士生 博士生 讲　师 教　授	043 043 043 043	基于本体的设计知识检索研究	中国机械工程	2007.18.21	
77	黄卫东 谢　强 丁秋林	博士生 讲　师 教　授	043 043 043	基于动态调度和任务委派的多型号产品并行研发能力管理	武汉大学学报(工学版)	2007.40.01	
78	张　磊 谢　强 丁秋林	博士生 讲　师 教　授	043 043 043	基于映射的异构信息源查询处理	计算机工程	2007.33.13	
79	张　磊 谢　强 李领治 丁秋林	博士生 讲　师 博士生 教　授	043 043 043 043	基于语义的Ontology聚类研究	应用科学学报	2007.25.01	
80	王有远 黄卫东 谢　强 丁秋林	博士生 博士生 讲　师 教　授	043 043 043 043	产品设计链管理及其关键技术研究	中国机械工程	2007.18.13	

信息科学与技术学院2007年发表论文目录-3.6

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
81	王有远 王　云 冯雪飞 丁秋林	博士生 博士生 博士生 教　授	043 043 043 043	协同产品设计链管理模式下的产品开发	科学学研究	2007.25.01	
82	曹永忠 李　斌 丁秋林	博士生 博士生 教　授	043 043 043	动态工作流的进化策略	微电子学与计算机	2007.24.09	
83	曹永忠 丁秋林	博士生 教　授	043 043	一种基于DFA的服务行为模型及其应用	微计算机信息	2007.23.7-3	
84	曹永忠 丁秋林 李　斌	博士生 教　授 博士生	043 043 043	基于有限自动机的Web服务行为的描述与发	现代电子技术	2007.00.02	
85	施化吉 丁秋林	博士生 教　授	043 043	一种基于MDCT变换的组合压缩音频水印算法	微计算机信息	2007.23.8-3	
86	施化吉 尹纪军 李星毅 丁秋林	博士生 硕士生 教授 教　授	043 043 外 043	利用影响因子遗传算法优化前馈神经网络	计算机应用研究	2007.24.11	
87	翟建军 乔新宇 丁秋林	博士生 教　授	043 043	基于扫掠法的六面体网络生成算法及实现	南京航空航天大学学报	2007.39.01	
88	杨淑群 丁树良 丁秋林	博士生 教　授	043 外 043	基于形式概念分析的状态空间图及其在CAT中的应用	武汉大学学报(工学版)	2007.40.04	
89	周　勇 朱梧槚	讲　师 教　授	043 043	多agent逻辑程序及其在协议验证中的应用	小型微型计算机系统	2007.28.01	
90	周　勇 朱梧槚	讲　师 教　授	043 043	一种验证非否认协议的新方法	电子与信息学报	2007.29.10	
91	杜庆伟 燕雪峰	讲　师 副教授	043 043	基于被动复制方式的容错CORBA模型	南京航空航天大学学报	2007.39.05	
92	刘学军	讲　师	043	Including probe-level uncertainty in model-based gene expression clustering	BMC Bioinformatics	2007.08.98	
93	曹子宁	副教授	043	A Spatial Logical Characterisation of Context Bisimulation*	Lecture Notes in Computer Science	2007.4435	
94	曹子宁	副教授	043	Towards an Epistemic Logic for Uncertain Agents*	Lecture Notes in Computer Science	2007.4696	
95	曹子宁	副教授	043	Bisimulations for a Distributed Higher order π-Calculus	Lecture Notes in Computer Science	2007.4711	
96	曹子宁	副教授	043	Axiomatising a Linear Higher Order π-Calculus	2007年ACST会议交流		
97	屠　莉 陈　崚	博士生 教授	043 外	Adaptive clustering lagorithm by ant's optimization	Journal of System Science and Information	2007.05.12	

信息科学与技术学院2007年发表论文目录-3.7

序号	姓名	职称	单位	论文题目	刊物、会议名称	年、卷、期	类别
98	刘 维 陈 峻	博士生 教授	043 外	A Fast Longest Common Subsequence Algorith for Biosequences Alignment	2007年Proceedings for the 1st International Conference on Computer and Computing Technology in Agriculture会议交流		
99	刘 维 陈 峻	博士生 教授	043 外	An efficient algorithm for multiple sequence alignment based on ant colony optimization and divide-and-conquer method	2007年Proceedings for the 1st International Conference on Computer and Computing Technology in Agriculture会议交流		
100	刘 维 陈 峻	博士生 教授	043 外	A Parallel Longest Common Subsequence Algorithm for Biosequences Alignment	2007年Proceedings for The Second International Conference on Scalable Information Systems会议交流		

第 12 卷 第 11 期
2007 年 11 月

中国图象图形学报
Journal of Image and Graphics

Vol. 12, No. 11
Nov., 2007

一种改进的基于密度的抽样聚类算法

胡彩平 秦小麟

（南京航空航天大学信息科学与技术学院，南京 210016）

摘 要 基于密度的聚类算法 DBSCAN 是一种有效的空间聚类算法，它能够发现任意形状的聚类并且有效地处理噪声。然而，DBSCAN 算法也有一些缺点，例如，①在聚类时只考虑空间属性没有考虑非空间属性；②在对大规模空间数据库进行聚类分析时需要较大的内存支持和 I/O 消耗。为此，在分析 DBSCAN 算法不足的基础上，提出了一种改进的基于密度的抽样聚类（improved density-based spatial clustering algorithm with sampling，IDBSCAS）算法，使之能够有效地处理大规模空间数据库，并且它不仅考虑了空间属性也考虑了非空间属性。2 维空间数据的测试结果表明，该算法是可行、有效的。

关键词 空间数据挖掘 空间聚类 密度 种子 非空间属性

中图法分类号：TP301.6 文献标识码：A 文章编号：1006-8961(2007)11-2031-06

An Improved Density-based Spatial Clustering Algorithm with Sampling

HU Cai-ping, QIN Xiao-lin

(*College of Information Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing* 210016)

Abstract DBSCAN is one of the effective spatial clustering algorithms, which can discover clusters of any arbitrary shape and handle the noise effectively. However, it has also several disadvantages. First, it is based on only spatial attributes without considering non-spatial attributes in the databases. Second, when DBSCAN handles large-scale spatial databases, it requires large volume of memory support and I/O cost. In this paper, an improved density-based spatial clustering algorithm with sampling(IDBSCAS) is developed, which not only clusters large-scale spatial databases effectively, but also considers spatial attributes and non-spatial attributes. Experimental results of 2-D spatial datasets show that the new algorithm is feasible and efficient.

Keywords spatial data mining, spatial clustering, density, seeds, non-spatial attributes

1 引 言

近年来，数据挖掘的研究已从关系型和事务型数据库扩展到空间数据库。由于雷达、红外、光电、卫星、电视摄像、电子显微成像、CT 成像等各种宏观与微观传感器的使用，空间数据的数量、大小和复杂性都在飞快地增长，已经远远超出了人们的解释能力。终端用户不可能详细地分析所有的这些数据，并提取感兴趣的空间知识，致使“空间数据爆炸，但知识贫乏”。因此，利用空间数据挖掘（spatial data mining，SDM）从空间数据库中自动或半自动地挖掘事先未知却潜在有用的空间模式已变得十分必要。所谓空间数据挖掘指的是从空间数据库中抽取隐含的知识、空间关系或非显式地存储在空间数据库中有意义的特征或模式。

聚类分析是数据挖掘领域中的一项重要的研究课题。所谓聚类，就是根据相似性对数据对象进行分组，通过发现空间数据的分布特征，使得每一个聚类中的数据有非常高的相似性，而不同聚类中的数据尽可能不同。迄今为止，人们已经提出了许多聚类算法。聚类分析算法通常有以下 5 类[1]：

基金项目：国家自然科学基金项目(60673127)；江苏省自然科学基金项目(BK2001045)

收稿日期：2006-04-13；改回日期：2006-08-22

第一作者简介：胡彩平(1977～)，男。博士研究生。研究方向为空间数据挖掘等。E-mail：hucaiping@nuaa.edu.cn

(1)基于划分的方法[2],如 K-means 算法、K-medoids 算法、PAM 算法、CLARA 算法、CLARANS 算法和 EM 算法;(2)基于层次的方法[3,4],如 CURE 算法和 BIRCH 算法;(3)基于密度的方法[5~7],如 DBSCAN 算法、OPTICS 算法、DENCLUE 算法和 GDBSCAN 算法;(4)基于网格的方法[8,9],如 STING 算法、CLIQUE 算法和 WaveCluster 算法;(5)基于模型的方法,如 COBWEB 算法。其中,DBSCAN 算法是一种颇具代表性的基于密度的空间聚类分析算法,它具有可以发现任意形状的聚类和有效屏蔽噪声数据干扰的优点。本文在分析 DBSCAN 算法不足的基础上,提出了一种改进的基于密度的抽样聚类算法(improved density-based spatial clustering algorithm with sampling,IDBSCAS)。该算法是通过减少区域查询的次数来降低聚类时间和 I/O 消耗。另外,由于本文算法不仅考虑了空间属性,还考虑了非空间属性,从而提高了聚类的质量。

2 关于 DBSCAN 算法

2.1 DBSCAN 算法的思想

DBSCAN 算法是第 1 个基于密度的空间聚类算法,其基本思想是:对于一个聚类中的每一个对象,在其给定半径 ε 的邻域中包含的对象不能少于某一给定的最小数目 $N_{\min}$,然后对具有密度连接特性的对象进行聚类。

对于对象集 $\boldsymbol{D}$,给定距离函数 d,参数 $\varepsilon \geqslant 0$ 和 $N_{\min} \geqslant 0$。下面是基于密度的聚类算法 DBSCAN 的一些定义[5]。

定义 1(ε-邻域) 给定 ε 和对象 $p \in \boldsymbol{D}$,以对象 p 为中心,以 ε 为半径的区域是对象 p 的 ε-邻域,用 $\boldsymbol{N}_\varepsilon(p)$ 来表示,即 $\boldsymbol{N}_\varepsilon(p) = \{q \in \boldsymbol{D} \mid d(p,q) \leqslant \varepsilon\}$。

定义 2(核心对象) 给定 ε 和 $N_{\min}$,若对象 p 的 ε-邻域 $\boldsymbol{N}_\varepsilon(p)$ 包含的对象个数 $|\boldsymbol{N}_\varepsilon(p)| \geqslant N_{\min}$,则称 p 为核心对象。

定义 3(直接密度可达) 给定 ε 和 $N_{\min}$,则对象 q 是从对象 p 出发直接密度可达的,当

(1) $q \in \boldsymbol{N}_\varepsilon(p)$

(2) $|\boldsymbol{N}_\varepsilon(p)| \geqslant N_{\min}$

定义 4(密度可达) 给定对象集合 $\boldsymbol{D}$,当存在一个对象链 $p_1, p_2, \cdots, p_n; p_1 = q, p_n = p$,对 $\forall p_i \in \boldsymbol{D}(i = 1,2,\cdots,n)$,且 p_{i+1} 是从 p_i 关于 ε 和 $N_{\min}$ 直接密度可达的,则称对象 p 从对象 q 关于 ε 和 $N_{\min}$ 密度可达(非对称)。

定义 5(密度相连) 如果对象集合 $\boldsymbol{D}$ 中存在一个对象 o,并使得对象 p 和 q 是从 o 关于 ε 和 $N_{\min}$ 密度可达的,那么对象 p 和 q 关于 ε 和 $N_{\min}$ 密度相连(对称)。

定义 6(聚类和噪声) 基于密度可达性最大的密度相连对象的集合称为聚类,不在任何聚类中的对象被认为是噪声。

为了发现一个聚类,DBSCAN 算法先开始从数据库中任意选取一个对象 p 开始执行一个区域查询,即寻找 p 的邻域对象,如果这个邻域的对象个数小于给定的阈值,则 p 为噪声,否则,p 邻域的所有对象形成一个聚类;然后这个邻域中的每个对象也执行同样的操作,同时看这些邻域中的对象是否能加到这个聚类中,重复执行同样的操作,直到这个聚类不能扩大为止。如果这个聚类不能被进一步扩大,那么 DBSCAN 算法就另外选取一个没有被标为某个聚类或噪声的对象,重复执行同样的操作。这个操作被重复执行,直到在数据库中的所有对象都被标为某个聚类或噪声。对于一个有 n 个对象的空间数据库,由于要执行 n 次区域查询,所以该算法的时间复杂性为 $O(n^2)$;在空间索引(如 R^*-树)的支持下[10],其复杂性为 $O(n\log n)$。

2.2 DBSCAN 算法的局限性及改进

实际上,DBSCAN 算法进行聚类的过程就是一个不断执行区域查询的过程,聚类过程的大部分时间都用在区域查询操作上。DBSCAN 算法对核心对象的邻域中包含的所有对象都执行区域查询来扩展聚类,显然这样时间花费很高。假设对聚类 $\boldsymbol{C}$ 中的某一给定核心对象 p 来说,可以想象它的邻域中所包含的所有对象的邻域将会互相覆盖。假定 q 是 p 邻域中的一个对象,如果它的邻域被 p 邻域中的其他对象的邻域所覆盖,则表明对 q 的区域查询是可以省掉的。这是因为 q 的邻域中所包含的对象可以通过对覆盖它的其他对象执行区域查询得到。也就是说,q 没有必要作为种子对象用于聚类扩展。实际上,对于稠密的数据集来说,在一个核心对象的邻域中有相当多的对象可以不用作为聚类扩展用的种子对象。这样从加速 DBSCAN 算法来讲,应当选择核心对象邻域中的部分代表对象,而不是像 DBSCAN 算法那样选择所有对象,作为种子对象用于聚类的扩展。这样就可以减少区域查询的次数和提高查询效率。

另外,DBSCAN 算法没有考虑非空间属性,如果非空间属性在聚类中起一个很大的作用的话,那么它就不适用,因为空间数据库不仅保存了空间物体的空间属性,也保存了它的非空间属性。空间属性表现为空间物体的方位信息,非空间属性表现为一个空间物体的其他属性,例如在一个地理信息系统中,一个社区的失业率就是非空间属性,社区的方位是空间属性。为了使聚类具有更高的质量,在聚类时不仅要考虑空间属性,也要考虑非空间属性。

3　一种改进的基于密度的抽样聚类算法

本文提出了一种改进的基于密度的抽样聚类算法。为了能够处理非空间属性,本文利用了纯度的概念,纯度是指邻域内相同非空间属性的对象个数和邻域内所有对象个数的比值。

3.1　基本概念

对于对象集 $\boldsymbol{D}$,给定距离函数 $dist$、参数 $\varepsilon \geqslant 0$、$MinPts \geqslant 0$ 和 $MinPur \geqslant 0$,用 $attr$ 来表示对象的一个非空间属性(很容易推广到多个非空间属性)。下面是改进的基于密度的抽样聚类算法的一些定义。

定义 7(ε-匹配邻域)　给定 ε 和对象 $p \in \boldsymbol{D}$,对象 p 的匹配邻域用 $\hat{N}_\varepsilon(p)$ 来表示,它被定义为 $\hat{N}_\varepsilon(p) = \{q \in \boldsymbol{D} \mid dist(p,q) \leqslant \varepsilon \text{ and } p.attr = q.attr\}$。

定义 8(纯核心对象)　给定 ε,$MinPts$,$MinPur$,若对象 p 的 ε-匹配邻域 $\hat{N}_\varepsilon(p)$ 包含的对象个数 $|\hat{N}_\varepsilon(p)| \geqslant MinPts$ 并且 $|\hat{N}_\varepsilon(p)|/|N_\varepsilon(p)| \geqslant MinPur$,则称 p 为纯核心对象。

定义 9(纯直接密度可达)　给定 ε,$MinPts$ 和 $MinPur$,如果满足以下的两个条件,则对象 q 是从对象 p 出发纯直接密度可达的,可用 $p \Rightarrow q$ 来表示。

(1) $q \in \hat{N}_\varepsilon(p)$

(2) $|\hat{N}_\varepsilon(p)| \geqslant MinPts$ 并且 $|\hat{N}_\varepsilon(p)|/|N_\varepsilon(p)| \geqslant MinPur$

定义 10(纯密度相连)　在对象集 $\boldsymbol{D}$ 中,当给定 ε,$MinPts$ 和 $MinPur$,则纯密度相连关系⇔满足如下两个条件:

(1) $\forall p \in \boldsymbol{D}, p \Leftrightarrow p$

(2) 当存在一个对象链 $p_1, p_2, \cdots, p_n; p_1 = q, p_n = p$,对 $\forall p_i \in \boldsymbol{D}(i = 1,2,\cdots,n)$,如果 $\forall p_{i+1} \Rightarrow p_i$ 或 $p_i \Rightarrow p_{i+1}$,则对象 $p \Leftrightarrow q$

从这个定义,就能推出如下的定理。

定理　在对象集 $\boldsymbol{D}$ 中,对于给定 ε,$MinPts$ 和 $MinPur$ 的纯密度相连关系⇔是一个等价关系。

理由很明显,证明略。

由于纯密度相连关系⇔是一个等价关系,所以纯密度相连关系⇔能在对象集 $\boldsymbol{D}$ 上确定它的一个划分,每一个划分就是对象集 $\boldsymbol{D}$ 的一个聚类或噪声。

定义 11(聚类)　在对象集 $\boldsymbol{D}$ 中的一个非空子集 $\boldsymbol{C}$ 是一个聚类,它必须满足以下两个条件:

(1) $\forall p, q \in \boldsymbol{D}$,如果 $p \in \boldsymbol{C}$ 并且 $p \Leftrightarrow q$,则 $q \in \boldsymbol{C}$

(2) $\forall p, q \in \boldsymbol{C}, p \Leftrightarrow q$

3.2　种子对象的选取

种子对象的选取非常重要,种子对象不能太多,亦不能太少。若太多,就难以发挥算法的效率,反之,如果太少,则种子对象邻域难以比较完全地覆盖其他对象的邻域,从而造成对象丢失,并影响到聚类质量和效率。所以有两个问题需要解决:(1)种子对象应该选多少;(2)如何选择种子对象。

在这里仅考虑 2 维空间数据,但这个方法也很容易推广到 2 维以上的空间数据。对于给定的 ε 和 $MinPts$,先假设对象 p 是核心对象,然后以对象 p 为中心,并以 ε 为半径画圆(如图 1 所示)。在这个圆上,先以对象 p 为原点画坐标系交圆周于 A、B、C 和 D 4 个点,然后画两条分别与 x 轴成 45°和 135°角的两条直径交圆周于 E、F、G 和 H 4 个点。本文把这些点称为代表边界对象(representative boundary objects, RBO)。在每一个象限中有 3 个 RBO,例如在右上象限中有 B、E 和 C,在右下象限中有 A、H 和 B。

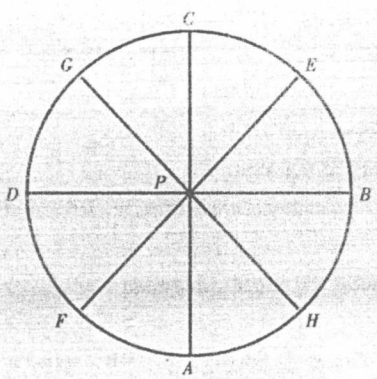


图 1　种子的选取

Fig. 1　Selection of seeds

下面给出了一种从核心对象的邻域中选择种子对象的算法。其基本思想是:对每一个 RBO,可在

它所在的象限内确定与它最近的对象，并把这个对象选为种子，即使一个种子对象到两个以上的 RBO 都是最近的，它也只被选一次。因此在 2 维空间范围内对任一对象的被选种子数最多是 8 个。一般情况下，对 d 维空间，由于有 3^d-1 个 RBO 和 2^d 个象限，因此被选种子数最多为 3^d-1 个，该算法的伪码如下：

```
Finding_seeds(Object,ε,Seeds)
  Initialize Seeds to NULL;
  Identify the RBOs about Object and ε;
  for each RBO do
    MinDist = ∞ ;
  Identify the Quadrant in which the RBO falls;
    for each unclassified object p in the Quadrant do
      distance = dist(RBO,p);
      if(distance≤ minDist)
        minDist = distance;
        minObject = p;
      endfor
    Seeds = Seeds∪{minObject};
  endfor
end
```

3.3 算法描述

算法名称：一种改进的基于密度的抽样聚类算法——IDBSCAS 算法（an improved density-based spatial clustering algorithm with sampling）。

输入：2 维平面上的空间数据库（DB），邻域半径 $\varepsilon\geq 0$，最小对象个数阈值 $MinPts\geq 0$，最小纯度阈值 $MinPur\geq 0$.

输出：聚类 Cluster.

算法伪码如下：

```
Algorithm IDBSCAS(DB,ε,MinPts,MinPur)
  ClusterId = nextId(NOISE);
  for each object o in DB do
    if o.ClId = UNCLASSIFIED
        if Expanding_cluster(o, ε, DB, MinPts, MinPur,
ClusterId)
        ClusterId = nextId(ClusterId);
  end
Expanding_cluster (Object, ε, DB, MinPts, MinPur,
ClId):Boolean
  Neighbours = DB.matchingNeighbours(Object,ε);
  if (Neighbours.size < MinPts) or
    (Neighbours.pur < MinPur)
      //Object 为噪声
      DB.changClId(Object,NOISE);
      return false;
  else//Object 为纯核心对象
    DB.changClId(Neighbours,ClId);
    Finding_Seeds(Object,ε,Seeds);
    Seed_list = Seed_list.add(Seeds);
    While Seed_list≠NULL do
    currentP = Seed_list.first();
      Neighbours = DB.matchingNeighbours(currentP,ε);
      if(Neighbours.size ≥ MinPts) and (Neighbours.pur ≥
        MinPur)
      //currentP 为纯核心对象
  Finding_Seeds(currentP,ε,Seeds);
    Seed_list = Seed_list.add(Seeds);
      for each object p in Neighbours do
        if p.ClId = UNCLASSIFIED or NOISE
DB.changClId(p,ClId);
        else
          ClId = p.ClId;//把两个纯密度相连的聚类合并
成一个聚类
      Seed_list.delete(currentP);
        endif
      endwhile
    return true;
  endif
end
```

在 IDBSCAS 算法中，当第 1 个纯核心对象找到后，就可通过 Finding_seeds 函数找出它所对应的种子对象作为聚类扩展用。在随后的聚类扩展中，新种子不断增加到种子集合 Seed_list 中，用于后续聚类的扩展。如此循环执行下去，直到 Seed_list 为空，这表明该聚类扩展完毕；然后 IDBSCAS 算法就另外选取一个没有被标为某个聚类或噪声的对象，重复执行同样的操作。这个操作被重复执行，直到在数据库中的所有对象被标为某个聚类或噪声。与 DBSCAN 算法相比，IDBSCAS 算法主要在以下 3 个方面进行了改进：(1) 可以处理非空间属性；(2) 在 Expanding_cluster 函数中，增加了 ClId = p.ClId 语句，通过它可以把两个纯密度相连的聚类进行合并；(3) 在 Expanding_cluster 函数中加入了 Finding_seeds 函数，用它在纯核心对象邻域中寻找种子对象执行聚类扩展，不需要再对邻域中的每个对象都执行区域查询，这样就大大节省了执行时间。

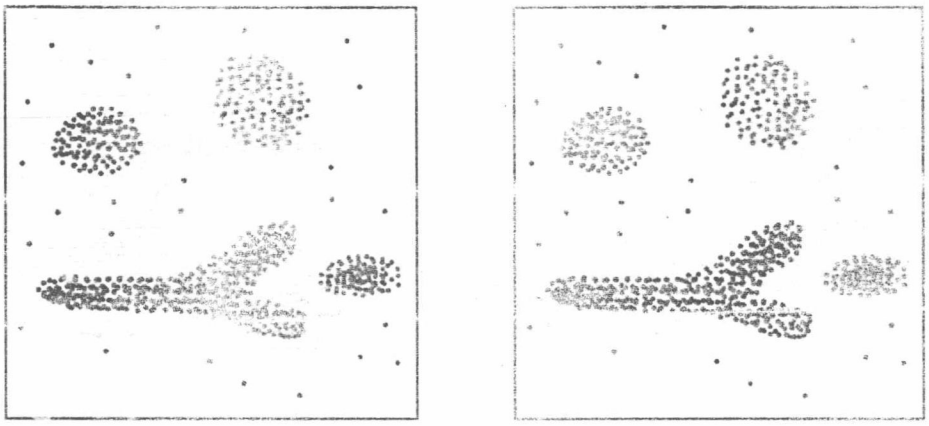
4 实验与分析

本文测试所用的数据为模拟数据和 DBSCAN

算法所使用的测试数据集,测试的硬件环境是:P4-1.5GHz的CPU,主存为256M,硬盘为40G。软件环境是:操作系统为 Microsoft Windows 2000 Professional,算法实现的语言为 C++。

4.1 聚类效果的比较

为了比较聚类的效果,本实验采用了文献[5]中的有203个点的数据集作为测试数据(如图2(a)所示)。在这个数据集中,用不同的颜色表示非空间属性。在其中的"Y"形点集中,它左右两部分的颜色不同,也就是说,它的非空间属性不同。IDBSCAS算法成功地确定了它的5个聚类和噪声,聚类的准确率为100%(如图2(a)所示)。但DBSCAN算法由于没有考虑非空间属性,它把"Y"点集确定为一个聚类,所以它得到的是4个聚类和噪声(如图2(b)所示)。可见,DBSCAN算法聚类的准确率明显比IDBSCAS算法聚类的准确率低。



(a) IDBSCAS 算法的聚类结果　　(b) DBSCAN 算法的聚类结果

图2　测试数据集

Fig. 2　Testing datasets

4.2 执行效率的比较

为了比较IDBSCAS算法和DBSCAN算法的执行效率,本文实验采用了一个对象个数从20 000到100 000个的模拟数据集。从图3可以看出,IDBSCAS算法的效率明显高于DBSCAN算法。在空间索引如R*-树的支持下,IDBSCAS算法的时间复杂性和DBSCAN算法一样,也为$O(n\log n)$。但IDBSCAS算法在执行区域查询的时候采用了抽样技术,由于不需要对核心对象邻域中的每个对象都执行区域查询,所以它比DBSCAN算法节省了不少时间。虽然IDBSCAS算法比DBSCAN算法多用了一个函数Finding_seeds,但并没有全面提高IDBSCAS算法的时间复杂性。假设邻域中对象个数为m,对象的维数为d,则函数Finding_seeds的时间复杂性为$O(md)$,但邻域中对象个数m和对象的维数d与数据库中对象的个数n相比是非常小的。

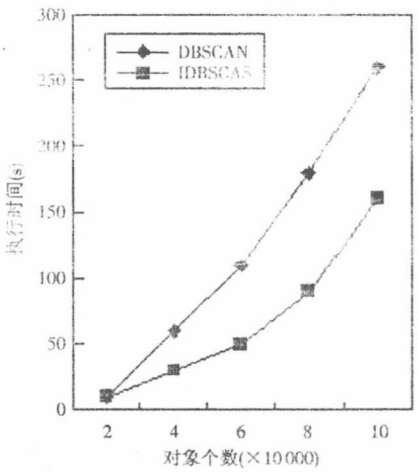


图3　性能比较

Fig. 3　Performance comparison

5 结　论

本文提出了一种改进的基于密度的抽样聚类算法,该算法保持了基于密度聚类算法可以发现任意形状的聚类和对噪声不敏感的优点,而且由于该算法采用了抽样技术,从而使聚类速度有了显著的提高;另外,由于该算法通过引入纯度的概念,在聚类时,不仅考虑了空间属性,而且也考虑了非空间属

性,从而提高了聚类的质量。

由于空间数据不断积累,空间数据库的规模不断扩大,因此如何提高聚类算法的执行效率,改进聚类结果的质量,仍然是有待进一步研究的内容。

参考文献(References)

1 Han Jiawei, Kamber Micheline(Fan Ming, Meng Xiao-feng, *et al* Translate). Data Mining Concepts and Techniques[M]. Beijing: China Machine Press, 2001.[Han Jiawei, Kamber Micheline(范明,孟小峰等译). 数据挖掘概念和技术[M]. 北京:机械工业出版社, 2001.]

2 Ng R T, Han Jiawei. CLARANS: A method for clustering objects for spatial data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, **14**(5): 1003 ~ 1016.

3 Guha S, Rastogi R, Shim K. CURE: An efficient clustering algorithm for large databases[A]. In: Proceedings of the ACM SIGMOD International Conference on Management of Data[C], Seattle, WA, USA, 1998: 73 ~ 84.

4 Zhang T, Ramakrishna R, Livny M. BIRCH: An efficient data clustering method for very large databases[A]. In: Proceedings of the ACM SIGMOD International Conference on Management of Data [C], Montreal, Canada, 1996: 103 ~ 114.

5 Ester M, Kriegel H, Sander J, *et al*. A density-based algorithm for discovering clusters in large dpatial databases with noise[A]. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining[C], Portland, Maine, USA, 1996: 226 ~ 231.

6 Ankerst M, Breunig M, Kriegel H, *et al*. OPTICS: Ordering points to identify the clustering structure[A]. In: Proceedings of the ACM SIGMOD International Conference on Management of Data[C], Philadelphia, PA, USA, 1999: 49 ~ 60.

7 Sander J, Ester M, Kriegel H, *et al*. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications[J]. Data Mining and Knowledge Discovery, 1998, **2**(2): 169 ~ 194.

8 Wang W, Yang J, Muntz R. STING: An statistical information grid approach to spatial data mining[A]. In: Proceedings of the 23th International Conference on Very Large Data Bases[C], Athens, Greece, 1997: 186 ~ 195.

9 Sheikholeslami G, Chatterjee S, Zhang A. WaveCluster: A multi resolution clustering approach for very large spatial databases[A]. In: Proceedings of the 24th International Conference on Very Large Data Bases[C], New York, USA, 1998: 428 ~ 439.

10 Beckmann N, Kriegel H P, Schneider R, *et al*. The R*-Tree: An efficient and robust access method for points and rectangles[A]. In: Proceedings of the ACM SIGMOD International Conference on Management of Data [C], Atlantic City, NJ, USA, 1990: 322 ~ 331.

空间数据挖掘研究综述 *)

胡彩平　秦小麟
(南京航空航天大学信息科学与技术学院　南京 210016)

摘　要　信息化的发展使得更多的空间数据被使用,因此获取空间知识也就越来越重要和有意义,并使得空间数据挖掘成为一个很有前途的研究领域。本文系统概括了空间分类和预测、空间聚类、空间孤立点和空间关联规则4类空间数据挖掘方法及其进展,最后探讨了空间数据挖掘的未来发展方向。

关键词　空间数据挖掘,空间分类和预测,空间聚类,空间孤立点,空间关联规则

A Survey of Spatial Data Mining Research

HU Cai-Ping　QIN Xiao-Lin
(College of Information Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 210016)

Abstract　More and more spatial data are used with the development of the information, therefore, obtaining the spatial knowledge becomes more and more important and meaningful, this makes spatial data mining become a promising research filed. In this paper, the proceedings of four methods used in spatial data mining, namely spatial classification and prediction, spatial clustering, spatial outlier, spatial association rules are systematically summarized. Finally, the future directions of spatial data mining are discussed.

Keywords　Spatial data mining, Spatial classification and prediction, Spatial clustering, Spatial outlier, Spatial association rules

1　引言

空间数据挖掘(Spatial Data Mining,SDM)指的是从空间数据库中抽取隐含的知识、空间关系或非显式地存储在空间数据库中的其它模式等[1]。空间数据挖掘需要综合数据挖掘(Data Mining,DM)与空间数据库技术,可用于对空间数据的理解,空间关系和空间与非空间数据间关系的发现、空间知识库的构造、空间数据库的重组和空间查询的优化等[2]。空间数据挖掘在地理信息系统、遥感、图像数据库探测、医学图像处理、导航、交通控制、环境研究以及许多使用空间数据的领域中有广泛的应用。

空间数据与传统数据相比,具有自己独特的特点。表现在两个方面:一是空间数据都与某一对象(地点)相关,空间数据中除包含以文字、字符为特征的非空间信息(属性信息)外,还含有以拓扑关系、距离关系、方向关系为特征的空间信息;二是空间数据具有空间自相关性[3],即每一个事物都与其它事物相关,但邻近事物间的相关性比距离较远的事物间的相关性要大得多。这就使得空间数据挖掘比传统数据挖掘更为困难,因此研发高效的空间数据挖掘技术是当前空间数据挖掘面临的主要挑战。

1989年8月美国底特律市召开的第一届国际联合人工智能学术会议上,从事数据库、人工智能、数理统计和可视化等技术的学者们,首次提出从数据库中发现知识(Knowledge Discovery in Database,KDD),标志着数据挖掘技术的诞生。空间数据挖掘在数据挖掘技术发展与海量空间数据积累的推动下,国内外都开展了积极的研究。加拿大西蒙法拉色大学计算机科学系的韩家炜教授领导的研究小组,就是对空间数据挖掘进行系统全面的研究[4~10]。1994年,在加拿大渥太华举行的GIS国际学术会议上,我国学者李德仁教授提出了从GIS数据库中发现知识的概念,系统分析了空间知识发现的特点和方法,认为它能够把GIS有限的数据变成无限的知识,并进一步用于精练和更新GIS数据,使GIS成为智能化的信息系统[11]。经过10多年的发展,国内外对空间数据挖掘已取得了一些成果,但空间数据挖掘的研究无论是在理论研究还是相关软件原型的研制等方面都还处于不断的探索。

2　空间数据挖掘方法

空间数据挖掘是计算机技术、数据库应用技术和管理决策支持技术等发展到一定阶段多学科交叉的新兴边缘学科,汇集了来自机器学习、模式识别、数据库、统计学、人工智能以及管理信息系统等各学科的成果[1]。一般地,空间数据挖掘可以分成4类挖掘方法[12]:空间分类和预测、空间聚类、空间孤立点和空间关联规则。

2.1　空间分类和预测

空间分类是指分析空间对象导出与一定空间特征有关的分类模式[12],如地区、高速公路或河流的领域;预测是根据某空间维找出变化趋势。分类和预测具有很大的相似性,在数据挖掘界广泛接受的观点是[2]:用预测法预测类标号为分类,用预测法预测连续值为预测。Ester等人[13]最早提出了一种空间对象分类方法,该方法采用ID3算法,并使用邻域图的概念,分类标准基于分类对象的非空间属性以及描述分类对象与其邻近位置相关对象间空间关系的属性、谓词和函数。该

*)航空科学基金项目(02F52033)资助;江苏省高技术研究计划项目(BG2004005)资助。**胡彩平**　博士研究生,研究方向:空间数据挖掘等;**秦小麟**　教授,博士生导师,研究方向:空间数据库、空间数据挖掘和安全数据库等。

方法的缺点是没有分析邻近对象非空间属性的聚合值。另外,该算法也没有进行相关性分析,可能会生成低质量的决策树。而且,该算法没有考虑非空间和空间属性值中可能存在的概念层次。

顾及决策树邻近对象的非空间属性的聚合值,基于分类对象的非空间属性,描述被分类对象和邻近特征的空间关系的属性、谓词和函数,Koperski 和 Han 提出了空间数据的两步决策分类法[9]。在查找样本对象的粗略描述后,利用机器学习的 Relief 算法提取空间谓词,合并空间谓词和非空间谓词为分类决策知识。但是基于决策树的分类算法不适合处理带有不完整信息的问题。空间数据分类标准中包含数据间的空间关系,从某个训练数据集来讲,空间属性极有可能缺失。如果输入数据出现了不一致、噪声等情况,决策树算法可能会造成误分,就会严重影响决策树算法的预测准确度,因而采用决策树空间分类算法不能很好地体现地理空间关系对于分类的影响。石云等人提出的基于 Rough Set 的空间数据分类方法[14],采用 Rough Set 方法进行空间对象分类,能够较好地反映空间和非空间数据之间的关系,为利用邻近区域中基于非空间属性的聚合值来对空间对象进行分类提供了可行性,较好地解决了上述问题。

文[15]提出了用空间自相关模型和马尔科夫随机域两种方法进行空间分类和预测,并且从理论和实验两个方面对这两种方法进行了比较。文[16,17]提出了一个框架,使用图相似度去预测沼泽地中鸟巢的位置。文[18]针对径向基函数不适合用于空间数据预测的特点,提出了分别在输入层、中间层和输出层中加入空间信息来进行空间数据预测。

2.2 空间聚类

聚类分析是数据挖掘领域中的一项重要的研究课题。所谓聚类[2],就是根据相似性对数据对象进行分组,发现空间数据的分布特征,使得每一个聚类中的数据有非常高的相似性,而不同聚类中的数据尽可能不同。迄今为止,人们已经提出了许多聚类算法。一般地,主要的聚类算法可以分为 5 类[2]:

(1)划分方法(Partitioning Method)。给定一个 n 个对象或元组的数据库,给定要构建的划分的数目 $k(k<n)$,划分方法首先创建一个初始划分。然后采用一种迭代重定位技术,尝试通过对象在划分间移动来改进划分。一个划分方法构建数据的 k 个划分,每个划分表示一个聚类。典型的划分算法如 k-means 算法、k-medoids 算法和 CLARANS 算法等。

k-means 算法[19]以 k 为参数,把 n 个对象分为 k 个聚类,以使聚类内具有较高的相似度,而聚类间的相似度较低。相似度的计算根据一个聚类的平均值(被看作聚类的重心)来进行。但 k-means 算法对孤立点是敏感的。k-medoids[20]算法不采用聚类中对象的平均值作为参照点,选用聚类中位置最中心的对象,即中心点。仍然是基于最小化所有对象与其参照点之间的相异度之和的原则来执行的。

CLARANS 算法[21]由 Ng 和 Han 提出,其聚类过程可以表示为查找一个图,图中的每个节点都是潜在的解决方案。在替换一个中心点后获得的聚类称为当前聚类的邻居。随意测试的邻居的数目由参数 maxneighbor 限制。如果找到一个更好的邻居,将中心点移至邻居节点,重新开始上述过程,否则在当前的聚类中生成一个局部最优。找到一个局部最优后,再任意选择一个新的节点,重新寻找新的局部最优。局部最优的数目被参数 numlocal 限制。可以看到,CLARANS 算法并不搜索遍所有的求解空间,也不限制在任何具体的采样中。CLARANS 算法每次迭代的计算复杂度与对象的数量基本呈线性关系。基于 CLARANS 算法的空间数据聚类算法也有两种:空间支配算法和非空间支配算法。CLARANS 方法的缺点是要求欲聚类的对象必须预先都调入内存里,这对非常大的空间数据库是不合理的。

(2)层次的方法(Hierarchical Method)。层次的方法对给定数据对象集合进行层次的分解,它分为凝聚层次聚类与分裂层次聚类。凝聚层次聚类是自底向上的策略,首先将每一个对象作为一个聚类,然后合并它们,直到满足某个条件;分裂层次聚类正好相反,首先把所有的对象看作一个聚类,然后逐渐细分成越来越小的聚类,直至达到某个终结条件为止。著名的层次方法有 BIRCH 算法和 CURE 算法等。

Zhang 等人提出了平衡迭代消减聚类算法 BIRCH[22],它是一种较为灵活的增量式聚类方法,能根据内存的配置大小自动调整程序对内存的需要。它有两个重要概念:聚类特征(Clustering Feature)和聚类特征树(CF-tree),它们用于概括聚类描述。聚类特征(CF)是一个三元组,给出对象子聚类的信息的汇总描述。给定某个子聚类中有 N 个 d 维的点或对象 $\{o_i\}$,则这个子聚类的聚类特征可表示为 $CF=(N, \overrightarrow{LS}, SS)$。其中,$N$ 是对象的个数,$\overrightarrow{LS}$ 是 N 个对象的线性和,即 $\overrightarrow{LS}=\sum_{i=1}^{N}\vec{o_i}$,它代表了这个子聚类的重心;$SS$ 是 N 个对象的平方和,即 $SS=\sum_{i=1}^{N}\vec{o_i}^2$,它代表了这个子聚类的直径大小,$SS$ 越小,这个子聚类聚得越紧。聚类特征树是一个满足两个条件的平衡树。两个条件分别是:分枝因子和子聚类直径的限制。分枝因子规定了树的每个节点的子女的最多个数;而子聚类直径体现了对一子聚类的直径大小的限制,即聚类特征的 SS 不能太大,否则不能聚为一类。非叶子节点上存储了它的子女的聚类特征的和,因此该节点总结了其子女的信息。

CURE 算法[23]选择基于质心和基于代表对象方法之间的中间策略。它不用单个质心或对象来代表一个子聚类,而是选择数据空间中固定数目的具有代表性的对象。

(3)基于密度的方法。其主要思想是:只要邻近区域的密度(对象的数目)超过某个阈值,就继续聚类。代表性算法有 DBSCAN 算法、OPTICS 算法、GDBSCAN 算法、DBRS 算法和 DENCLUE 算法等。

DBSCAN 算法[24]是第一个基于密度的空间聚类算法,用来发现带有噪声的空间数据库中任意形状的聚类。该算法的效率较高,但算法执行前需输入阈值参数。为了解决这个难题,提出了 OPTICS 算法[25]。它没有显示地产生一个子聚类的集合,它为自动和交互的聚类分析计算一个聚类次序,这个次序代表了数据的基于密度的聚类结构。它包含的信息等同于从一个宽广的参数设置范围所获得的基于密度的聚类。GDBSCAN 算法[26]是 DBSCAN 算法的推广,它不仅能对点进行聚类,也能对线或多边形进行聚类。但 DBSCAN 算法仅能对点进行聚类,在现实生活中,要聚类的空间对象都用点来抽象有时不可行。DBSCAN 算法进行聚类的过程就是一个不断执行区域查询的过程,聚类过程的大部分时间都用在区域查询操作上。DBSCAN 算法对核心对象的邻域中包含的所有对象都执行区域查询来扩展聚类,显然没有必要。DBRS 算法[27]仅选择核心对象邻域中的部分代表对象,而不是像 DBSCAN 算法那样选择所有对象,作为种子对象用于聚类的扩展。这样就可以减少区域查询的次数,提高查询效率。另