PaperFree检测报告简明打印版

相似度：17.11%
编　号：FRYHNATUZWFWMOSL
标　题：面向高维数据的PCA-Hub聚类方法
作　者：-
长　度：52726字符
时　间：2017-04-11 10:02:15
比对库：中国学位论文全文数据库；中国学术期刊数据库；中国重要会议论文全文数据库；英文论文全文数据库；互联网资源；自建比对库

相似资源列表(学术期刊，学位论文，会议论文，英文论文等本地数据库资源)
1. 相似度：0.23%　篇名：《约束变密度界面反演方法》
来源：《地球物理学进展》　年份：2013　作者：张盛
2. 相似度：0.18%　篇名：《数据压缩算法在煤矿类软件系统中的应用研究》
来源：《工矿自动化》　年份：2013　作者：张卫国
3. 相似度：0.17%　篇名：《慈善，为什么如此重要？》
来源：《社会工作》　年份：2011　作者：高峰
4. 相似度：0.14%　篇名：《改进的k—means聚类算法在客户细分中的应用研究》
来源：《河北经贸大学学报》　年份：2014　作者：杜巍
5. 相似度：0.14%　篇名：《混合高斯模型运动检测算法优化》
来源：《计算机应用研究》　年份：2013　作者：胥欣
6. 相似度：0.13%　篇名：《中原城市群第三产业发展水平评价》
来源：《当代经济》　年份：2014　作者：郭凌霄
7. 相似度：0.11%　篇名：《中国海洋经济与国民经济周期的协动性分析》
来源：《海洋经济》　年份：2013　作者：何佳霖
8. 相似度：0.11%　篇名：《基于领域本体的文本资料聚类算法改进研究》
来源：《情报科学》　年份：2013　作者：龚光明
9. 相似度：0.11%　篇名：《混合高斯模型的混合em算法研究及聚类应用》
来源：《新疆大学硕士论文》　年份：2015　作者：曹红丽
10. 相似度：0.09%　篇名：《复杂制造业集配中心的货物采运数学模型及其解析》
来源：《企业经济》　年份：2014　作者：吴益伟
11. 相似度：0.09%　篇名：《k-means聚类算法在负荷曲线分类中的应用》
来源：《电力系统保护与控制》　年份：2011　作者：刘莉
12. 相似度：0.08%　篇名：《中国城镇居民收入分布的变迁研究》
来源：《吉林大学社会科学学报》　年份：2013　作者：孙巍
13. 相似度：0.08%　篇名：《k-均值聚类算法初始中心选取相关问题的研究》
来源：《湖南大学硕士论文》　年份：2008　作者：吴晓蓉
14. 相似度：0.07%　篇名：《基于路径的划分聚类算法研究》
来源：《西安理工大学硕士论文》　年份：2010　作者：王赛芳
15. 相似度：0.07%　篇名：《一种基于加性噪声的通用隐写分析算法》
来源：《电子测量与仪器学报》　年份：2012　作者：叶学义
16. 相似度：0.06%　篇名：《初始化中心点优化的K-means算法》
来源：《科技信息》　年份：2011　作者：周杨
17. 相似度：0.06%　篇名：《聚类分析在客户细分领域中的研究与应用》
来源：《华北电力大学（河北）硕士论文》　年份：2015　作者：朵春红
18. 相似度：0.06%　篇名：《一种选取初始聚类中心的方法》
来源：《计算机工程与应用》　年份：2004　作者：刘立平
19. 相似度：0.06%　篇名：《新型背景混合高斯模型》
来源：《中国图象图形学报》　年份：2011　作者：白向峰
20. 相似度：0.06%　篇名：《不确定数据的高效聚类算法》
来源：《广西师范大学学报：自然科学版》　年份：2011　作者：李云飞

21. 相似度：0.06% 篇名：《一种基于改进BFS算法的主题搜索技术研究》
来源：《现代图书情报技术》 年份：2013 作者：乔建忠
22. 相似度：0.06% 篇名：《图像分割与合成方法的研究》
来源：《天津大学硕士论文》 年份：2007 作者：徐高奎
23. 相似度：0.05% 篇名：《感谢我的医生老师》
来源：《糖尿病之友》 年份：2011 作者：赵凤琴
24. 相似度：0.05% 篇名：《基于AFS理论与PCA的体育教师评价体系构建方法研究》
来源：《廊坊师范学院学报：自然科学版》 年份：2014 作者：冯兴刚
25. 相似度：0.05% 篇名：《基于加权Voronoi图的民航机场空间服务范围研究》
来源：《交通运输系统工程与信息》 年份：2013 作者：冯社苗
26. 相似度：0.05% 篇名：《时空相关面板数据模型估计方法研究》
来源：《商业经济与管理》 年份：2011 作者：陈青青
27. 相似度：0.05% 篇名：《基于半监督学习的在线评论挖掘应用》
来源：《计算机光盘软件与应用》 年份：2012 作者：张建欣
28. 相似度：0.04% 篇名：《粒子群和遗传算法在农业工程中的应用》
来源：《农机化研究》 年份：2013 作者：尤文坚
29. 相似度：0.04% 篇名：《计算机数据加密技术的方法及应用探析》
来源：《美与时代：城市》 年份：2013 作者：王珂琦
30. 相似度：0.04% 篇名：《相似性度量在基因表达聚类分析中的应用研究》
来源：《现代电子技术》 年份：2012 作者：孙杰
31. 相似度：0.04% 篇名：《浅谈数据挖掘技术的概念》
来源：《科技视界》 年份：2013 作者：赵淑君
32. 相似度：0.04% 篇名：《一种基于区域划分的RSSI定位方法》
来源：《计算机系统应用》 年份：2014 作者：李锋
33. 相似度：0.03% 篇名：《医学专业英语论文撰写格式》
来源：《河南科技大学学报：医学版》 年份：2004 作者：左巨森
34. 相似度：0.03% 篇名：《面向信息特征模式识别的核方法研究综述》
来源：《现代情报》 年份：2014 作者：黄炜
35. 相似度：0.03% 篇名：《人脸识别技术研究》
来源：《电脑编程技巧与维护》 年份：2011 作者：张明超
36. 相似度：0.03% 篇名：《数据挖掘算法研究》
来源：《现代电子技术》 年份：2015 作者：周牒岚
37. 相似度：0.03% 篇名：《RAW格式到底是什么？》
来源：《影像视觉》 年份：2011 作者：Madder（翻译）
38. 相似度：0.03% 篇名：《面板数据的贝叶斯Lasso分位回归方法》
来源：《数量经济技术经济研究》 年份：2013 作者：李翰芳
39. 相似度：0.03% 篇名：《数据挖掘常用算法分析》
来源：《科技信息》 年份：2012 作者：张健
40. 相似度：0.03% 篇名：《关于女啊力口强金融学专业本科生统计学思维训练的思考》
来源：《中外企业家》 年份：2013 作者：魏峰
41. 相似度：0.03% 篇名：《大数据时代索引员的使命》
来源：《中国索引》 年份：2013 作者：朱晓霄
42. 相似度：0.03% 篇名：《基于PCA的SOM网络在基因数据聚类分析中的应用》
来源：《软件导刊》 年份：2015 作者：程国建
43. 相似度：0.03% 篇名：《构建数据仓库过程中的数据清洗研究》
来源：《图书与情报》 年份：2013 作者：刘喜文
44. 相似度：0.03% 篇名：《数据挖掘技术概述》
来源：《黑龙江科技信息》 年份：2012 作者：董欢
45. 相似度：0.03% 篇名：《网络近邻择优策略下的股市羊群行为演化模型及仿真》
来源：《中国管理科学》 年份：2013 作者：卞曰瑭
46. 相似度：0.03% 篇名：《基于知识发现的读者决策采购研究》
来源：《图书馆学研究》 年份：2014 作者：刘军
47. 相似度：0.03% 篇名：《组织冗余与绩效关系的国内外研究综述》
来源：《商业时代》 年份：2013 作者：武博
48. 相似度：0.03% 篇名：《毕业论文致谢词》

来源：《学习博览》 年份：2011 作者：卫小平
49. 相似度：0.03% 篇名：《仿生传感智能感官检测技术及其在云南特色农产品品质检测中的应用》
来源：《安徽农业科学》 年份：2013 作者：许文方
50. 相似度：0.02% 篇名：《大数据时代数字图书馆发展浅析》
来源：《江苏技术师范学院学报》 年份：2013 作者：李翠萍
51. 相似度：0.02% 篇名：《网络结构与银行系统性风险》
来源：《管理科学学报》 年份：2014 作者：隋聪
52. 相似度：0.02% 篇名：《自动气象站数据异常的处理》
来源：《黑龙江气象》 年份：2011 作者：杜红
53. 相似度：0.02% 篇名：《算法交易的市场影响研究》
来源：《管理科学学报》 年份：2014 作者：王宇超

相似资源列表(百度文库，豆丁文库，博客，新闻网站等互联网资源)

1. 相似度：1.22% 标题：《聚类分析Cluster analysis数学模型对冲基金方法_柳州文铮_新浪博客》
来源：http://blog.sina.com.cn/s/blog_7f6f656101018f4s.html
2. 相似度：0.97% 标题：《简单易学的机器学习算法——基于密度的聚类算法DBSCAN - null的...》
来源：http://blog.csdn.net/google19890102/article/details/37656733
3. 相似度：0.61% 标题：《【统计】Principal components analysis(PCA)主成分分析(更新!)_...》
来源：http://blog.sina.com.cn/s/blog_6e334fc90100zrs5.html
4. 相似度：0.50% 标题：《聚类分析_百度百科》
来源：http://baike.baidu.com/item/%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90
5. 相似度：0.48% 标题：《数据挖掘聚类算法 - zhoubl668的专栏:远帆,梦之帆! - 博客频道 - ...》
来源：http://blog.csdn.net/zhoubl668/article/details/6639801
6. 相似度：0.39% 标题：《基于密度的聚类算法(DBSCAN),density-based spatical clustiny of ...》
来源：http://www.dictall.com/indu/178/1772616D952.htm
7. 相似度：0.38% 标题：《聚类算法_百度百科》
来源：http://baike.baidu.com/item/%E8%81%9A%E7%B1%BB%E7%AE%97%E6%B3%95
8. 相似度：0.38% 标题：《聚类——层次聚类(Hierarchical Clustering) - 计算机科学与艺术- 博客...》
来源：http://blog.csdn.net/lanchunhui/article/details/50877161
9. 相似度：0.37% 标题：《子空间聚类算法的研究及应用_图文_百度文库》
来源：http://wenku.baidu.com/view/a706fc34a32d7375a417802c.html
10. 相似度：0.30% 标题：《PCA主成分分析- JustForCS - 博客园》
来源：http://www.cnblogs.com/JustForCS/p/5295266.html
11. 相似度：0.30% 标题：《基于集成学习的H-K聚类算法研究_图文_百度文库》
来源：http://wenku.baidu.com/view/f08d0e8ef7ec4afe05a1df09.html
12. 相似度：0.29% 标题：《半监督组合分类算法研究与应用_图文_百度文库》
来源：http://wenku.baidu.com/view/922ed492aa00b52acec7ca32.html
13. 相似度：0.27% 标题：《【机器学习】K-means聚类算法初探 - the Pensieve - 博客 ...》
来源：http://blog.csdn.net/skyline0623/article/details/8154911
14. 相似度：0.25% 标题：《知识发掘是什么意思《德语助手》德汉-汉德词典_知识发掘的德语...》
来源：http://www.godic.net/dicts/de/%E7%9F%A5%E8%AF%86%E5%8F%91%E6%8E%98
15. 相似度：0.25% 标题：《1聚类分析介绍 - 仰望星空的麦兜 - 博客频道 - CSDN.NET》
来源：http://m.blog.csdn.net/article/details?id=48712041
16. 相似度：0.22% 标题：《层次聚类(Hierarchical Clustering)_我爱吃糖不长蛀牙_新浪博客》
来源：http://blog.sina.com.cn/s/blog_816101080102wm2v.html
17. 相似度：0.21% 标题：《聚类(2)——层次聚类Hierarchical Clustering - 姜文晖的博客- 博客...》
来源：http://blog.csdn.net/jwh_bupt/article/details/7685809
18. 相似度：0.21% 标题：《聚类——混合高斯模型 Gaussian Mixture Model - jerrice - ...》
来源：http://www.cnblogs.com/jerrice/p/4313592.html
19. 相似度：0.19% 标题：《非常感谢我的家人,是她们的无私支持和悉心照顾,我才能专心 ...》
来源：http://www.39394.com/fanwen/doc/257214_3.html
20. 相似度：0.19% 标题：《【机器学习】聚类算法:层次聚类 - ZhangPY的专栏 - 博客频道 - ...》
来源：http://blog.csdn.net/lg1259156776/article/details/52127232
21. 相似度：0.18% 标题：《聚类技术及其在银行客户细分中的应用研究_图文_百度文库》
来源：http://wenku.baidu.com/view/6748d0ff4693daef5ef73deb.html

22. 相似度：0.18% 标题：《数据挖掘聚类算法的分析和应用研究_百度文库》
来源：http://wenku.baidu.com/view/b17ab8e8fab069dc50220123.html
23. 相似度：0.18% 标题：《常见的数据分布（正态分布，ZIPF分布，偏态分布）_佳佳hi ...》
来源：http://blog.sina.com.cn/s/blog_5caa94a001014x42.html
24. 相似度：0.16% 标题：《curse of dimensionality维数灾难 - Orisun - 博客园》
来源：http://www.cnblogs.com/zhangchaoyang/articles/2801525.html
25. 相似度：0.15% 标题：《混合高斯模型（Mixtures of Gaussians）和EM算法 - AndyJee - 博客园》
来源：http://www.cnblogs.com/AndyJee/p/3732766.html
26. 相似度：0.14% 标题：《高频交易策略 - 道客巴巴》
来源：http://www.doc88.com/p-3985541027572.html
27. 相似度：0.14% 标题：《数据挖掘中聚类算法 - bigshuai - 博客园》
来源：http://www.cnblogs.com/bigshuai/articles/2599504.html
28. 相似度：0.14% 标题：《从K近邻算法、距离度量谈到KD树、SIFT+BBF算法- July_ - 博客园》
来源：http://www.cnblogs.com/v-July-v/archive/2012/11/20/3125419.html
29. 相似度：0.14% 标题：《聚类分析-概念 - 搜狗百科》
来源：
http://baike.sogou.com/v363477.htm?fromTitle=%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90
30. 相似度：0.14% 标题：《聚类分析：经理人分享百科》
来源：http://www.managershare.com/wiki/%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90
31. 相似度：0.13% 标题：《如何区分人工智能、机器学习和深度学习?》
来源：http://m.sohu.com/n/483653693/
32. 相似度：0.13% 标题：《数据聚类_互动百科》
来源：http://www.baike.com/wiki/%E6%95%B0%E6%8D%AE%E8%81%9A%E7%B1%BB
33. 相似度：0.13% 标题：《峰度和偏度原码_清淨菩提心_新浪博客》
来源：http://blog.sina.com.cn/s/blog_02cf67f00102v06t.html
34. 相似度：0.12% 标题：《机器学习_百度百科》
来源：
https://wapbaike.baidu.com/item/%e6%9c%ba%e5%99%a8%e5%ad%a6%e4%b9%a0/217599?adapt=
35. 相似度：0.12% 标题：《维数灾难- 必应》
来源：http://www.bing.com/knows/search?FORM=BKACAI&mkt=zh-
cn&q=%E7%BB%B4%E6%95%B0%E7%81%BE%E9%9A%BE
36. 相似度：0.11% 标题：《聚类评价指标 Rand Index,RI,Recall,Precision,F1_精品文库_IThao...》
来源：http://www.ithao123.cn/content-2319454.html
37. 相似度：0.11% 标题：《聚类- 知识小屋 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/zd0303/article/details/8425563
38. 相似度：0.10% 标题：《【机器学习】机器学习(十二、十三):K-means算法、高斯....》
来源：http://m.blog.csdn.net/article/details?id=46564237
39. 相似度：0.10% 标题：《高维数据的聚类析方法研究及其应用.pdf文档全文免费阅读、在线看》
来源：http://max.book118.com/html/2015/1231/32362006.shtm
40. 相似度：0.09% 标题：《视觉机器学习--K-means算法- 有梦放飞- 博客园》
来源：http://www.cnblogs.com/xiaotongtt/p/6182835.html
41. 相似度：0.09% 标题：《基于密度方法的聚类_图文_百度文库》
来源：
http://wenku.baidu.com/link?url=4FfcsYuL4WSJzECaf8kKOe8VLUckGbiQlvisKF40sT1s4hOq8WePlXJ9jw
y2JkseyWf1gJK4IgmP4YjdSuaXHhuwWecMVpZAZJG-rou
42. 相似度：0.09% 标题：《决策树算法介绍及应用 - 文章 - 伯乐在线》
来源：http://blog.jobbole.com/89072/
43. 相似度：0.09% 标题：《基于潜在语义索引的文本聚类算法研究_图文_百度文库》
来源：http://wenku.baidu.com/view/5a55ade8f8c75fbfc77db255.html
44. 相似度：0.08% 标题：《K近邻算法- pi9nc的专栏- 博客频道- CSDN.NET》
来源：http://blog.csdn.net/pi9nc/article/details/9068437
45. 相似度：0.08% 标题：《轮廓系数_百度百科》
来源：
http://baike.baidu.com/item/%E8%BD%AE%E5%BB%93%E7%B3%BB%E6%95%B0/17361607?fr=aladd
46. 相似度：0.07% 标题：《数据挖掘聚类方法的研究 - 我还是我的日志 - 网易博客》

来源：http://blog.163.com/crazyzcs@126/blog/static/129742050200991110584546 3/
47. 相似度：0.07% 标题：《聚类- u014665416的博客 - 博客频道 - CSDN.NET》
来源：http://m.blog.csdn.net/article/details?id=51900510
48. 相似度：0.07% 标题：《15DBSCAN》
来源：http://3y.uu456.com/bp_157bq3qc6l4i6jo0x0ij_1.html
49. 相似度：0.06% 标题：《Spark 机器学习《一》 - skynumone的专栏- 博客频道- CSDN.NET》
来源：http://blog.csdn.net/skynumone/article/details/42913457
50. 相似度：0.06% 标题：《NP难问题求解综述_happy芥末青豆_新浪博客》
来源：http://blog.sina.com.cn/s/blog_702cefb50101gene.html
51. 相似度：0.06% 标题：《基于网格的聚类算法分析与研究_基于网格的聚类算法..._爱问共享资料》
来源：http://ishare.iask.sina.com.cn/f/19969072.html
52. 相似度：0.05% 标题：《聚类分析_百度百科》
来源：https://wapbaike.baidu.com/item/%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90
53. 相似度：0.05% 标题：《CRM中的模糊C均值(FCM)客户聚类算法研究_CNKI学问》
来源：http://xuewen.cnki.net/CJFD-HEBG200402031.html
54. 相似度：0.05% 标题：《[转载]高斯混合模型的期望最大化算法_记得曾经_新浪博客》
来源：http://blog.sina.com.cn/s/blog_ad2d83330102vhl3.html
55. 相似度：0.05% 标题：《基于模糊聚类分析的脑MRI图像分割算法的研究_百度文库》
来源：http://wenku.baidu.com/view/cfd62102a2161479171128fd.html
56. 相似度：0.05% 标题：《聚类算法实现与分析- liuheng0111的博客- 博客频道- CSDN.NET》
来源：http://blog.csdn.net/liuheng0111/article/details/52349006
57. 相似度：0.04% 标题：《聚类分析的方法及应用 - 马海祥博客-专注于分享seo思维和 ...》
来源：http://www.mahaixiang.cn/sjfx/746.html
58. 相似度：0.04% 标题：《Letters Using batch algorithm for kernel blind source separation $_...》
来源：http://www.nexoncn.com/read/f19d02020159220cc0b0f1f7.html
59. 相似度：0.04% 标题：《关于聚类的相关总结_宋兵乙_新浪博客》
来源：http://blog.sina.cn/dpool/blog/s/blog_7103b28a0102w81n.html
60. 相似度：0.04% 标题：《关于子空间聚类subspace clustering – 【人人分享-人人网】》
来源：http://blog.renren.com/share/112881531/10392469901
61. 相似度：0.04% 标题：《一篇文章透彻解读聚类分析及案例实操(一)_浪迹孤独_新浪博客》
来源：http://blog.sina.com.cn/s/blog_13eacc8800102wybk.html
62. 相似度：0.04% 标题：《硕士毕业论文致谢精选》
来源：http://www.cnrencai.com/lunwen/lunwenzhixie/617842.html
63. 相似度：0.04% 标题：《基于K-Means的文本聚类 - 阿罗的技术博客 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/freesum/article/details/7376006
64. 相似度：0.04% 标题：《基于密度的优化初始聚类中心K-means算法研究-AET-电子技术应用》
来源：http://www.chinaaet.com/article/3000015218
65. 相似度：0.04% 标题：《聚类算法在银行客户细分中的研究和应用_图文_百度文库》
来源：http://wenku.baidu.com/view/58719a47fe4733687e21aa72.html
66. 相似度：0.04% 标题：《matlab学习：人脸识别之PCA（ Principal Component ...》
来源：http://www.cnblogs.com/yingying0907/archive/2012/11/18/2775584.html
67. 相似度：0.04% 标题：《weighted Kernel k-means 加权核k均值算法理解及其实现(一)-爱编程》
来源：https://www.w2bc.com/article/206072
68. 相似度：0.04% 标题：《研究生论文致谢词》
来源：http://www.cdfds.com/lunwenzhixie/13209.html
69. 相似度：0.04% 标题：《致谢语岁月如梭,如歌 转眼间,三年的研究生求学生涯即将结束,站在...》
来源：https://www.douban.com/note/354791823/
70. 相似度：0.03% 标题：《机器学习中的维数灾难 - Forever-守望 - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/zbc1090549839/article/details/38929215
71. 相似度：0.03% 标题：《各种聚类算法及改进算法的研究》
来源：http://www.360doc.com/content/10/1229/10/4232513_82242690.shtml
72. 相似度：0.03% 标题：《[转]基于GPU的K-Means聚类算法_Constantine_新浪博客》
来源：http://blog.sina.com.cn/s/blog_64b7bad301012fbp.html
73. 相似度：0.03% 标题：《亚洲最大的网络棋牌:信誉保证2017优秀毕业论文致谢词范文》
来源：http://www.nblixue.com.cn/zhixie/637469.html

74. 相似度：0.03% 标题：《模糊C均值聚类算法的实现- liu_xiao_cheng的专栏- 博客频道- CSDN....》
来源：http://blog.csdn.net/liu_xiao_cheng/article/details/50471981
75. 相似度：0.03% 标题：《非参数边际距离最大化准则及其应用.pdf 全文免费在线阅读-淘豆网》
来源：http://www.taodocs.com/p-18122346-2.html
76. 相似度：0.03% 标题：《PCA原理分析和Matlab实现方法(三) - guyuealian的博客 - 博客频道 - ...》
来源：http://blog.csdn.net/guyuealian/article/details/68487833
77. 相似度：0.03% 标题：《在论文完成之际,我要特别感谢我的指导老师XX老师的热情关怀和悉...》
来源：http://wenwen.sogou.com/z/q66153347.htm
78. 相似度：0.03% 标题：《K-means算法原理 - Little_Rookie - 博客园》
来源：http://www.cnblogs.com/nxld/p/6376496.html
79. 相似度：0.03% 标题：《Pearson,Kendall和Spearman三种相关分析方法的异同_xu1003_新浪...》
来源：http://blog.sina.com.cn/s/blog_548d137e0101874n.html
80. 相似度：0.03% 标题：《算法&模型 - 我和我追逐的梦~~~ - 博客频道 - CSDN.NET》
来源：http://blog.csdn.net/heyongluoyao8/article/details/48494401
81. 相似度：0.03% 标题：《关于聚类与Mapreduce--梦飞翔的地方(梦翔天空)》
来源：http://www.dreamflier.net/blog/user1/3/2598.html
82. 相似度：0.03% 标题：《核聚类与支持向量聚类- bluenight专栏- 博客频道- CSDN.NET》
来源：http://blog.csdn.net/chl033/article/details/4758592
83. 相似度：0.03% 标题：《基于PCA和半监督聚类的入侵检测算法研究_图文_百度文库》
来源：http://wenku.baidu.com/view/cb996a8076eeaeaad1f330b2.html
84. 相似度：0.03% 标题：《聚类集成理论与其在图像分类中的应用/罗会兰:图书比价:琅琅比价网》
来源：http://www.langlang.cc/3295276.htm
85. 相似度：0.03% 标题：《分层聚类 - 了凡春秋USTC | 格物致知，诚意正心》
来源：http://chunqiu.blog.ustc.edu.cn/?p=466
86. 相似度：0.03% 标题：《聚类_宋兵乙_新浪博客》
来源：http://blog.sina.com.cn/s/blog_7103b28a0102w4e1.html
87. 相似度：0.03% 标题：《聚类算法总结-tombaby-ChinaUnix博客》
来源：http://blog.chinaunix.net/uid-10289334-id-3758310.html
88. 相似度：0.03% 标题：《基于聚类的异常检测技术的研究_图文_百度文库》
来源：http://wenku.baidu.com/view/7c87c390dd88d0d233d46a74.html
89. 相似度：0.03% 标题：《数据挖掘(六)---聚类分析:其他问题与算法 - lemon的日志 - 网易博客》
来源：http://blog.163.com/zhoulili1987619@126/blog/static/353082012015295101874O/
90. 相似度：0.03% 标题：《WGCNA算法研究笔记》
来源：http://www.mamicode.com/info-detail-548009.html
91. 相似度：0.03% 标题：《【机器学习】聚类分析(一)——k-means算法- 软件开发其他- 红黑联盟》
来源：http://www.2cto.com/kf/201608/533053.html
92. 相似度：0.03% 标题：《聚类算法分析- u013593585的专栏- 博客频道- CSDN.NET》
来源：http://blog.csdn.net/u013593585/article/details/51534205
93. 相似度：0.03% 标题：《根据《关于“k-means算法在流式细胞仪中细胞分类的应用”的学习笔...》
来源：http://www.cnblogs.com/assiamein9/p/3669748.html
94. 相似度：0.02% 标题：《基于主动学习的K-Hub聚类算法--《计算机系统应用》2016年03期》
来源：http://www.cnki.com.cn/Article/CJFDTOTAL-XTYY201603032.htm
95. 相似度：0.02% 标题：《k-均值聚类_百度文库》
来源：http://wenku.baidu.com/link?url=pjwxoSB8YRpeftDVq--zRIhf1YQmzR-sE-rkgzjox0yvDFgvRtHTgRlvoJXe0QISajSj3j1WOLjmZAZD-4VuKdHM_cgMHlnYpKtxVQJIiou

全文简明报告

附 录

摘 要

{86% : 机器学习(Machine Learning)是一门人工智能科学, }{ 66% : 该领域的主要研究对象为人工智能,机器学习是通过机器自主学习的方式来处理人工智能中的问题,特别是如何在经验学习中改善具体算法的性能。 }近几十年机器学习在概率论、计算复杂性理论、统计学、逼近论等领域均有发展,已形成一门多领域交叉学科。机器学习通过设计和分析让机器可以自主“学习”的算法以便从海量数据中自动分析出有

价值的模式或规律,从而对未知数据进行预测。机器学习大致可以分为下面四种类别:{86%：监督学习(Supervised Learning)、无监督学习(Unsupervised Learning)、半监督学习(Semi-supervised Learning)以及增强学习(Reinforcement Learning)。 }{ 72%：机器学习已广泛应用于诸多领域:数据挖掘、计算机视觉、搜索引擎、自然语言处理、语音和手写识别、生物特征识别、DNA序列测序、医学诊断、检测信用卡欺诈和证券市场分析等。 }

{ 61%：聚类分析是机器学习中的一种无监督学习, }{ 69%：近些年来受到越来越多的关注。 }{100%：聚类分析(Cluster Analysis, }{ 67%：亦称为群集分析)是把相似的对象通过静态分类的方法分成不同的簇或子集, }{ 72%：使得在同一个簇中的对象都具有某些相似的属性。 }{ 63%：传统的聚类分析算法大致可分为以下五种:划分聚类(Partitioning Clustering)、层次聚类(Hierarchical Clustering)、基于密度的聚类(Density-Based Clustering)、基于网格的聚类(Grid-Based Clustering)和基于模型的聚类(Model-Based Clustering)。 }传统的聚类分析算法更倾向于在低维数据空间中进行聚类分析。{ 61%：然而由于现实生活中数据的复杂性和多样性, }传统聚类算法在处理诸多现实问题和任务时,往往不能科学地进行聚类分析,尤其在高维数据和海量数据上更是如此。这是因为在高维数据空间中利用传统聚类算法时常常会出现下述两个问题:第一,{ 55%：高维数据存在大量冗余、噪声的特征使得不可能在所有维中均存在簇;第二,高维空间中的数据分布十分稀疏, }其数据间的距离几乎相等。显然,{ 72%：基于距离的传统聚类方法无法在高维空间中基于距离来构建簇。 }这便是机器学习中令人头疼的维数灾难(Curse of Dimensionality)问题[1]。近年来,{ 66%：“维数灾难”已成为机器学习的一个重要研究方向。 }{ 59%：随着科技的发展使得数据获取变得愈加容易,而数据规模愈发庞大、复杂性越来越高,如海量Web文档、基因序列等, }其维数从数百到数千不等,甚至更高。{ 63%：高维数据分析虽然十分具有挑战性,但是它在信息安全、金融、市场分析、反恐等领域均有很广泛的应用。 }

为了解决维数灾难的问题,本文引入了hubness这一全新的概念,并在原有的hub聚类算法上通过实验分析后,对hub算法进行了改进。Hubness这一概念是在2010年由Milos˘ Radovanovic Ì•等人提出的[2],hubness描述的是这样一种现象:在k近邻列表中,某些点容易频繁地出现在其它点的k近邻列表中。在数据集中样本点出现在其它点的k近邻列表中的次数称为k-occurrences,随着维度的增加,k-occurrences的分布会逐渐向右倾斜,这将会导致hubs的出现。Hubs通常是指具有非常高的k-occurrences的样本点,换言之,hubs易于频繁地出现在其它点的k近邻类列表中。通过探究这种现象的根源,发现这是高维数据空间数据统计分布的一种内在属性[2]。Milos˘ Radovanovic Ì•等人利用这种内在属性对基于距离度量的各种机器学习方法进行了深入的研究,{ 66%：包括监督学习方法、半监督学习方法和无监督学习方法。 }在无监督学习方面,hub聚类分析算法有以下四种:deterministic、probabilistic、hybrid和kernel。{ 67%：这四种方法均为K-Means算法的扩展。 }{ 57%：Hub聚类算法虽然可以在高维数据空间中进行聚类分析, }但是它却忽略了高维数据空间中的冗余和噪声数据,从而无法获得更优的簇结构以及更快的聚类收敛速度。

本文针对hub聚类分析算法的上述问题,提出了一种PCA-Hub聚类分析算法用于解决高维数据空间中的冗余和噪声数据,以便获得更好的簇结构和更快的聚类收敛速度。PCA-Hub聚类算法是以k-occurrences的偏度与本征维数强烈正相关为理论基础,通过构建数据集的KNN邻域矩阵,以偏度的变化率作为降维依据选出理想的k个主成分,之后再对降维后的数据集进行聚类分析。实验结果表明,PCA-Hub聚类算法相比之前的聚类算法在轮廓系数上平均提高了15%;当数据集的维数或者k-occurrences的偏度较高时,PCA-Hub聚类算法对近邻数的选择不敏感;在实验环境和实验参数一致的情况下,PCA-Hub聚类算法的结果在很大程度上具有一致性。

PCA-Hub聚类算法虽然可以很好地解决高维数据空间中的冗余和噪声数据,{ 68%：然而随着数据集尺度和数据集维数的不断增加, }PCA-Hub聚类算法的耗时将会变得越来越严重甚至是不可接受。因此,本文提出了一种Quick PCA-Hub聚类分析算法分别从快速搜索k个主成分和快速搜索最近邻居两方面加快PCA-Hub算法的聚类分析速度。实验结果表明,Quick PCA-Hub聚类算法相比之前的聚类算法在轮廓系数上平均提高了8%;Quick PCA-Hub在高维数据空间中搜索理想的k个主成分时表现出了巨大的优势。

重庆大学硕士学位论文

中文摘要

关键词:Hub聚类,高维数据,偏度,本征维度,主成分分析

II

I

ABSTRACT

W:

重庆大学硕士学位论文

英文摘要

Key words: wind blades, anti-icing, hydrophobic coatings, frozen speed, wind tune test, natural icing

IV

III

重庆大学硕士学位论文

目 录

绪 论

{ 67% : 本章主要介绍论文研究背景及其意义, }阐述论文研究方向及其主要内容,同时对论文的整体结构作简要说明。

研究背景及意义

由于当今科学技术的发展越来越迅捷,并且云计算等新兴大数据处理技术也在计算机等诸多领域持续发展,因此人们对大型数据表现出前所未有的关注。信息网络的快速传播使得现实生活中数据几乎呈现出指数增长的趋势,随着网络数据的持续增加和网络数据结构的持续复杂,使得数据分析变得愈加困难。当今社会数据的过快产生使得我们身处在一个“被信息所淹没,但却渴望从中获取知识”的环境中[3]。对于这些大量、增长速度持续增加并且结构异常复杂的数据,{ 64% : 传统的数据处理方法已变得不再适用。 }于是,{ 56% : 一种基于大数据的处理方法应运而生。 }{ 62% : 数据挖掘的主要目标是从大量数据中提取出有价值的模式和知识, }然后将其转变为人类可理解的结构,以便后续的工作使用[4]。

在大型的数据集中,数据挖掘通过机器学习、人工智能、统计学等交叉方法从而发现有价值的模式和知识。数据挖掘的过程是对大型数据进行监督或半监督的分析,从而获得之前未知的有意义的潜在信息,例如数据的聚类(通过聚类分析)、数据的异常信息(通过利群点检测)和数据之间的联系(通过关联规则分析)。数据挖掘的对象的类型并无限制,可以使任意类型的数据,{ 60% : 不管是结构化的数据、半结构化的数据, }还是异构型的数据[5]。{ 56% : 数据挖掘的主要过程如图1.1所示。 }数据挖掘的过程通常定义为以下三大阶段:第一、预处理阶段:在获取到目标数据集后,{ 59% : 有必要对多变量数据进行分析, }处理那些包含噪声和含有缺失数据的观测量;第二、数据挖掘阶段:数据挖掘过程通常涉及六种常见的任务,异常检测(异常/变化/偏差检测)、关联规则学习(依赖建模)、聚类、分析、回归以及汇总,这些均是利用数据挖掘技术从原有的数据集中发现未知的有价值信息;第三、结果验证阶段:通常,{ 56% : 数据挖掘是有目的地挖掘未知的有价值信息,然而这些信息是否符合预期一般可以通过结果验证来实现。 }

{92% : 数据挖掘的方法包括监督式学习、无监督式学习、半监督学习以及增强学习。 }监督学习是从已知的训练数据集中获得某种函数用于预测未知的数据集。{ 55% : 监督学习训练集中的目标是人为标注的。 }{85% : 常见的监督式学习包括分类、估计、预测。 }{ 74% : 无监督学习与监督学习的不用之处在于训练集是没有人为标注的。 }{80% : 常见的无监督式学习包括聚类、关联规则分析。 }{100% : 半监督学习介于监督学习与无监督学习之间。 }增强学习是基于环境而行动,从而获得最大化的预期利益。{ 68% : 聚类分析是一种常见的无监督式学习, }{ 64% : “物以类聚,人以群分”,无论是自然科学还是现实世界中均有各种各样的分类问题。 }{ 55% : 在数据挖掘中,聚类分析是研究分类问题的一种数据分析方法。 }{ 57% : 聚类分析是把大量复杂的数据通过聚类器将其分成若干不同的类别或更多的子集, }换言之,聚类分析的目的是尽可能地增大簇内部的相似性同时减小簇之间的相似性。{ 78% : 聚类分析在诸多领域均有应用,包括机器学习、数据挖掘、模式识别、图像分析以及生物信息等。 }

图 1.1 数据挖掘逻辑图

Figure 1.1 Data mining logic diagram

然而随着科学技术的发展,数据集的获取愈加便捷,{ 69% : 同时数据集的维数也不断增加。 }虽然传统的聚类分析算法在低维数据空间可以获得良好的聚类效果,但是由于高维数据空间中数据的稀疏性以及距离集中等问题使得传统聚类算法难以进行科学的聚类分析。因此,{ 61% : 针对高维数据空间中的聚类分析变得很有意义。 }

国内外研究现状

{ 55% : 随着科学技术的发展,人们处理大型复杂数据的需求越来越强烈, }可以处理大型数据的数据挖掘在学术界也越来越受到关注。多年来,数据挖掘的相关理论不断完善和发展,而且其商业价值也逐步显现。{ 61% : 在数据挖掘中,聚类分析一直是其重要的组成部分, }自然也受到了研究者的高度关注。聚类分析是在1932年由两位人类学专家Driver和Kroeber首次提出的,1938年Zubin将其引入到了心理学领域。{ 62% : 就聚类分

析本身而言,它并不是一个具体的算法,}而是处理某一类问题的通用规则。不同的聚类器可以定义不同的簇结构以及搜寻不同的簇的规则。主流的簇概念包括簇内对象之间的最小距离、数据空间的密集区域以及间隔或者特定的分布。{ 59% : 因此,聚类分析可以被表示为自动化地实现多目标的优化问题。 }然而,{ 61% : 聚类分析本身并不是一个自动化的过程,}而是一个不断迭代的知识发现过程或是交互式多目标优化的过程。在这个迭代过程中需要不断修改数据预处理方式以及模型参数直到到达预期的结果。不同的数据集和不同的结果预期用途决定了聚类算法的选取和参数的设定(包括要使用的距离函数、密度阈值或者预期聚类的数量)。{ 66% : 当前主流的聚类分析算法可分为以下几类: }

①层次聚类算法

{ 73% : 层次聚类算法也称作基于连通性的聚类算法,}{ 58% : 其核心思想是若两个对象越接近,}那么它们的相关性就越强。这些算法基于对象间的距离将彼此连通从而形成不同的簇。在很大程度上,一个簇可以由该簇内的最大连通距离来表示。不同的距离会形成不同的簇,这可以通过树形结构来表示,这也是层次聚类名称的来源。通常而言,{ 73% : 层次聚类可分为两大类:自底向上(agglomerative)和自顶向下(divisive)。 }{ 69% : 自底向上是开始时所有数据点均各自为一个类别,然后每次迭代将距离最近的两个类合并,直到只有一个类为止。 }自顶向下的思想与自底向上的思想正好完全相反。{ 62% : 虽然层次聚类的核心思想比较简单,但是计算复杂度却比较高。 }因为上述的三种方法均需要计算所有点对之间的距离,{ 72% : 而且算法也表明每次迭代只能合并两个子类,这是非常耗时的。 }

②划分聚类算法

在基于中心的聚类中,由中心向量表示一个簇,该簇的中心元素不一定是数据集中的元素。当簇的个数为定值k时,K-Means聚类方法给出了关于最优化问题的形式定义:搜寻k个簇中心,然后将对象划分给与之最近的簇中心所在的簇。K-Means聚类算法的目标是最小化簇内平方和(WCSS within-cluster sum of squares)。{ 64% : 但是最优化问题本身是一个NP难问题,}{ 68% : 因此常见的方法是找到其近似解。 }最著名的近似方法是Lloyd's[6],也就是K-Means算法。然而,{ 60% : K-Means算法只能找到局部最优值,}而且使用的是随机的初始值。{ 70% : 虽然K-Means算法算法有很多变形,}但是大多数基于K-Means算法的变形算法的最大缺点之一是需要预先指定簇的个数k。此外,由于这些算法均是基于簇中心的,所以更容易形成大小相似的簇,这通常会导致簇之间不正确的边界切割。

③基于分布的聚类

与统计学最密切相关的聚类模型是基于分布的模型。对象的分布越相似,它们分配在同一个簇的可能性越大。{ 57% : 该方法易于处理类似于人工生成的数据集(从分布中随机取样)。 }{ 57% : 虽然基于分布的聚类方法有着优秀的理论基础,}但是它们却容易导致过拟合问题。因此,我们需要对这类模型添加复杂性约束。从理论上而言,{ 59% : 选择越复杂的模型越能更好地约束数据,}但是选择合适的复杂度模型却是十分困难的,{ 64% : 最常用的模型是高斯混合模型(使用期望最大化算法)。 }{ 55% : 高斯混合模型是以数据服从高斯混合分布为假设的,换言之,数据可以看作是从多个高斯分布随机选择出来的。 }{ 58% : 相对于K-Means算法,高斯混合模型每一次的迭代计算量均较大。 }由于高斯混合模型的聚类方法来源于EM算法,因此可能会产生局部极值,这与初始参数的选取密切相关。{ 64% : 高斯混合模型不仅可以用于聚类分析,同样也可用于概率密度估计。 }

④基于密度的聚类

{ 77% : 基于密度的聚类分析算法的主要目是搜寻被低密度区域分割的高密度区域。 }{ 65% : 不同于基于距离的聚类算法(基于距离的聚类算法是以数据集为球状簇为前提进行聚类的),基于密度的聚类算法可以发现任意形状的簇,这有利于处理带有噪声点的数据。 }{ 76% : 在基于密度的聚类分析算法中,}{ 64% : 分布在稀疏区域的对象通常被认为是噪声或边界点。 }目前,{83% : 最流行的基于密度的聚类算法是DBSCAN(Density-Based Spatial Clustering of Application with Noise)[7]。 }{ 60% : OPTICS[8]是DBSCAN的变形,}它并不需要为范围参数选择合适的值就可以产生与分层聚类相似的分层效果。Density-Link-Clustering结合单连通聚类和OPTICS的思想,完全消除了参数,并且通过使用R树索引增强了聚类性能。{ 77% : DBSCAN和OPTICS的主要缺点是它们是通过某种程度的密度下降来检测簇边界的。 }{ 55% : 此外,它们无法检测现实生活数据中普遍存在的内在簇结构。 }而DBSCAN的变形方法EnDBSCAN可以解决此类问题[9]。Mean-shift方法基于核密度估计[10],将每个对象移动到其附近最密集的区域。最终,对象会收敛到密度的局部最大值。但是由于昂贵的迭代过程和密度估计,Mean-shift通常比DBSCAN的效率要低。

图 1.2 聚类分析算法分类

Figure 1.2 The classification of clustering analysis

近年来,诸多学者投身于提高现有算法性能的研究中[12][13],{ 63% : 其中包括CLARANS(Ng和Han,1994)[14][18]和BIRCH(Zhang等,1996)[15]。 }由于处理海量数据集的需求日益增长,所以研究人员试图

通过换取聚类性能以增加所产生簇的语义分析能力,这一意愿引起了pre-clustering的发展,其中以canopy聚类最具代表性[16]。Canopy聚类算法可以处理超大型数据集,但是所得到的“聚类”仅仅是对数据集的粗略预分割,之后仍需使用现有的聚类分析算法对这些预分割数据集进行聚类。学者们一直在进行各种各样的聚类算法尝试,比如基于seed的聚类方法[17]。

本文研究的主要内容

{ 56% : 从聚类分析的国内外现状可以看出, }基于划分的聚类算法适用于凸集或者类球形的数据集;{ 64% : 基于密度的聚类算法依赖于数据集的密度分布, }{ 59% : 因此可以发现任意形状的簇,适用范围更加广泛。 }但是高维数据空间的稀疏性和距离集中使得无论是基于距离还是密度的聚类算法都变得不再适用。因此,本文引进了hubness这一概念,并利用逆近邻的偏度很好地解决了这一问题。针对传统的聚类分析算法所存在的问题,{ 79% : 本文主要完成了以下几方面的工作: }

{ 58% : 第一、研究了课题的相关背景及其意义, }从数据挖掘的实践应用和理论基础两方面对聚类分析的国内外现状进行了概述总结,着重阐述了聚类分析在高维数据中所遇到的挑战。

第二、通过查看相关的聚类分析算法文献,{ 56% : 例如层次聚类算法、划分聚类算法、基于分布的聚类算法以及基于密度的聚类算法等, }对以上算法有了一定的了解并总结出了算法各自的优缺点及其适用范围。

{ 67% : 第三、比较了现有的聚类分析算法, }{ 66% : 如K-Means,DBSCAN等, }并且针对其基本思想进行了分析研究。研究表明,无论是基于距离的聚类算法还是基于密度的聚类算法都无法解决维数灾难的问题,同时研究了可在高维数据空间聚类的hub聚类算法。

第四、将原有的hub聚类分析算法与主成分分析相结合,提出“无损”降维聚类的PCA-Hub聚类算法,该算法解决了高维数据中存在的冗余和噪声数据,{ 63% : 同时在不损失重要的有价值信息的情况下对数据集进行降维, }增强了算法的聚类性能。PCA-Hub聚类算法虽然可以很好地解决高维数据空间中的冗余和噪声数据,{ 68% : 然而随着数据集尺度和数据集维数的不断增加, }PCA-Hub聚类算法的耗时将会变得越来越严重甚至是不可接受。因此,本文提出了一种Quick PCA-Hub聚类分析算法分别从快速搜索k个主成分和快速搜索最近邻居两方面加快PCA-Hub算法的聚类分析速度。

第五、本文采用若干个UCI数据集,将PCA-Hub聚类算法与传统的K-Means算法和hub聚类算法进行对比实验,揭示了无论数据集是否呈现出较高的hubness情况下该算法均可以取得不错的聚类效果。若数据集未呈现出较高的hubness现象时,传统的K-Means方法更为适用;然而,当数据集表现出较高的hubness现象时,hub聚类算法则会取得不错的聚类效果。Quick PCA-Hub聚类算法相比之前的聚类算法在轮廓系数上平均提高了8%,在高维数据空间中搜索理想的k个主成分时相比PCA-Hub聚类算法表现出了巨大的优势。根据对比实验结果揭示了该算法的聚类效果相对较佳,同时给出了详细的实验结果分析,以及深入讨论了各算法的适用性和优缺点。

论文的章节排版

本文大致分为5大章节,详细的论文排版结构如下:

第1章 绪论:主要概述了数据挖掘领域的研究背景及其意义,着重分析了传统聚类算法和hub聚类算法,{ 59% : 同时阐述了聚类分析的现实意义和价值。 }

第2章 聚类分析概述:介绍了当前主流的几大经典聚类算法,着重研究了基于距离的划分聚类方法和基于密度的聚类分析方法,并且列出了各类算法的适用性和优缺点。最后,对当前聚类分析的发展趋势和热点问题进行了简要地概述和总结。

第3章 PCA-Hub聚类算法:详细地介绍了hubness这一现象的起源、定义以及hub聚类分析算法,{ 55% : 通过实验对hub聚类算法进行了深入的分析研究,并归纳出了其适用性及优缺点。 }由于hub聚类算法未处理高维数据空间中的冗余和噪声数据从而无法获得更优的簇结构以及更快的聚类收敛速度,因此本文提出了PCA-Hub的聚类算法用于解决此问题。PCA-Hub的聚类算法是以逆近邻偏度的变化率作为降维标准,在降维的同时保留了大量有价值信息从而提高了聚类效果。{ 60% : 在UCI的若干个数据集上进行实验分析, }将该算法与K-Means算法以及hub聚类算法等进行实验对比分析,并通过轮廓系数作为聚类结果的评价指标。实验结果表明,PCA-Hub聚类算法相比之前的聚类算法在轮廓系数上平均提高了15%;当数据集的维数或者k-occurrences的偏度较高时,PCA-Hub聚类算法对近邻数的选择不敏感;在实验环境和实验参数一致的情况下,PCA-Hub聚类算法的结果在很大程度上具有一致性。

第4张 Quick PCA-Hub聚类算法:PCA-Hub聚类算法虽然可以很好地解决高维数据空间中的冗余和噪声数据,{ 68% : 然而随着数据集尺度和数据集维数的不断增加, }PCA-Hub聚类算法的耗时将会变得越来越严重甚至是不可接受。因此,本文提出了一种Quick PCA-Hub聚类分析算法分别从快速搜索k个主成分和快速搜索最近邻居两方面加快PCA-Hub算法的聚类分析速度。实验结果表明,Quick PCA-Hub聚类算法相比之前的聚类算

法在轮廓系数上平均提高了8%;Quick PCA-Hub在高维数据空间中搜索理想的k个主成分时表现出了巨大的优势。

{ 64% : 第5章 总结与展望:主要对本文中的重点工作进行了概括总结,同时指出研究过程中存在的不足, }并且对将来的工作以及研究重点作了简要的说明。

1绪 论

1绪 论

49

22

25

聚类分析概述

{87% : 聚类分析(Cluster analysis),也称为群集分析, }常用于统计数据分析,在诸多领域均拥有广泛应用。{ 63% : 聚类是将元素分成不同的组别或者更多的子集, }使得分配到相同簇中的元素彼此之间比其它的数据点更为相似,也就是说,聚类算法的目的是要增加类内的相似性并减小类间的相似性。{ 67% : 聚类分析不同于分类,分类常被视为是有监督的学习, }{ 63% : 而聚类分析一般归纳为一种无监督式的学习。 }聚类分析不依赖于先验知识(类标签),只依赖自身属性,通过自身属性可以区分簇之间的相似性或者对象之间的相似性。{ 76% : 聚类分析作为一门十分有效的数据分析技术, }{ 71% : 常被应用于机器学习、数据挖掘、模式识别以及生物信息等领域。 }本章将详尽地描述各种聚类算法,根据归纳的文献综述,{ 59% : 聚类分析算法大致可以分为以下四种: }{ 64% : 层次聚类算法、基于中心的聚类算法、基于分布的聚类算法以及基于密度的聚类算法。 }{ 56% : 本章主要阐述了各类聚类算法的基本思想,并总结归纳出了各自的优缺点和适用范围。 }

聚类分析的定义

由于难以对簇的概念作出准确定义,从而导致有诸多的聚类算法产生[19]。虽然不同的聚类算法对簇的概念定义不同,但是它们却有却有一个共同之处:簇是一组数据对象。不同聚类分析算法使用了不同的聚类模型,掌握聚类模型是了解聚类分析算法的关键,下面仅列出了主流的聚类模型:

①连通性模型(Connectivity Models),例如层次聚类基于距离连通性构建模型;

②中心性模型(Centroid Models),例如K-Means算法将单个平均向量表示每个簇类;

③分布模型(Distribution Models),使用统计分布对聚类进行建模,例如由EM算法使用的是多变量正态分布;

④密度模型(Density Models),例如DBSCAN和OPTICS将簇定义为数据空间中的连接密集区域;

⑤子空间模型(Subspace Models),在Biclustering(也称为协同聚类或双模式聚类)中使用群集成员和相关属性建模;

⑥组模型(Group Models),一些算法不提供精确模型,仅提供分组信息;

⑦基于图论的模型(Graph-Based Models):图中的节点分成了若干个子集,在这些子集中的每两个点都通过一条边相连。

一般而言,{ 55% : 不同的聚类模型对应着不同的聚类分析算法。 }但是从本质上来说,聚类分析就是一组簇的集合,{ 58% : 该集合通常包含数据集的所有对象, }因此聚类分析可以大致地分为以下两种:

①硬聚类(hard clustering):每个数据对象要么属于一个簇,要么不属于任何簇;

②软聚类(soft clustering,也称为模糊聚类fuzzy clustering):每个数据对象有一定的概率属于每个簇。

如果需要进一步对聚类进行划分,可参照如下严格的聚类划分结果:

①严格划分聚类,每个对象正好属于一个簇;

{ 65% : ②包含离群点的严格划分聚类, }{ 55% : 对象也可以不属于任何簇,那么它将会被视为离群点; }

{ 57% : ③重叠聚类(也称作可替代聚类或多视图聚类),虽然通常是硬聚类,但对象也可能属于多个簇; }

④分层聚类:属于子集群的对象同时也属于父集群;

{ 57% : ⑤子空间聚类:在唯一定义的子空间内尽可能的没有重叠簇。 }

聚类分析算法是根据它们的聚类模型进行分类的,没有客观的“正确的”聚类算法,正如Vladimir已经指出的,“聚类是在旁观者的眼中(Clustering is in the eye of the beholder)”。针对特定的问题,除非有数据理论依据,否则需要通过实验进行选择合适的聚类算法[19]。

常用的聚类分析算法

下面仅列出最主流的聚类算法。

层次聚类算法

{ 73% : 层次聚类算法也称作基于连通性的聚类算法, }{ 58% : 其核心思想是若两个对象越接近, }那么它们的相关性就越强。这些算法基于对象间的距离将彼此连通从而形成不同的簇。在很大程度上,一个簇可以由该簇内的最大连通距离来表示。不同的距离会形成不同的簇,这可以通过树形结构来表示,这也是层次聚类名称的来源。{ 73% : 层次聚类可分为两大类:自底向上(agglomerative)和自顶向下(divisive)。 }{ 69% : 自底向上是开始时所有数据点均各自为一个类别,然后每次迭代将距离最近的两个类合并,直到只有一个类为止。 }自顶向下的思想与自底向上的思想正好完全相反。{ 57% : 计算两个类之间的距离共有三种方法: }

{ 69% : ①Single Linkage(也称作nearest-neighbor),是指类之间的距离为这两个类中距离最近的两个点之间的距离,然而这容易导致“Chaining”现象的发生。 }“Chaining”现象是指原本整体相距较远的簇只因其中个别点之间的距离较近而被合并,若依此合并最终会得到比较松散的簇;

{ 74% : ②Complete Linkage:是Single Linkage的反面极端,是指类之间的距离为这两个类中距离最远的两个点之间的距离。 }负面效果显而易见,原本已经很近的两个簇,只因有不配合的点存在而不能合并。

{97% : ③Group Average: }是指把类之间的距离定义为所有点对的距离的平均值。

{ 62% : 虽然层次聚类的核心思想比较简单,但是计算复杂度却比较高。 }因为上述的三种方法均需要计算所有点对之间的距离,{ 72% : 而且算法也表明每次迭代只能合并两个子类,这是非常耗时的。 }

基于中心的聚类算法

{ 57% : 在基于质心的聚类中,由中心向量表示一个簇, }该簇的中心元素不一定是数据集中的元素。当簇的个数为定值k时,K-Means聚类方法给出了关于最优化问题的形式定义:搜寻k个簇中心,然后将对象划分给与之最近的簇中心所在的簇。K-Means的目的是最小化簇内平方和(WCSS within-cluster sum of squares)。{ 68% : 最优化问题本身是一个NP难问题, }{ 68% : 因此常见的方法是找到其近似解。 }最著名的近似方法是Lloyd's[d5],也就是K-Means算法。然而,{ 56% : K-Means算法智能只能找到局部最优值, }而且使用的是随机的初始化值。{ 65% : K-Means的衍生算法作了如下的改进: }选择多次运行的最优值,并且将簇中心限定为数据集中的元素(K-Medoids);选择中值作为簇中心(K-Medians);较少的选择随机值作为簇中心(K-Means++);{ 70% : 或者可以模糊聚类(Fuzzy C-Means)。 }大多数基于K-Means算法的最大缺点之一是需要预先指定簇的个数k。此外,由于这些算法均是基于簇中心的,所以更容易形成大小相似的簇,这通常导致簇之间不正确的边界切割。

K-Means有以下重要的理论性质。首先,{ 58% : 它可以将数据划分为Voronoi图结构。 }其次,在理论上它与最近邻概念接近,因此在机器学习领域大受欢迎。{ 68% : 最后,它可以被视为基于模型分类的变形, }{ 61% : 并且K-Means算法是EM算法的一种变形。 }

基于分布的聚类算法

与统计学最密切相关的聚类模型是基于分布的模型。对象的分布越相似,它们分配在同一个簇中的可能性越大。{ 57% : 该方法易于处理类似于人工生成的数据集(从分布中随机取样)。 }{ 57% : 虽然基于分布的聚类方法有着优秀的理论基础, }但是它们却容易导致过拟合问题。因此,我们需要对这类模型添加复杂性约束。从理论上而言,{ 59% : 选择越复杂的模型越能更好地约束数据, }然而选择合适的复杂度模型却是十分困难的。最常用的基于分布的聚类算法的模型是高斯混合模型(使用期望最大化算法)。{ 55% : 高斯混合模型是以数据服从高斯混合分布为假设的,换言之,数据可以看作是从多个高斯分布随机选择出来的。 }{ 55% : 每个高斯混合模型由k个高斯分布组成,每个高斯分布被称作一个"Component", }高斯混合模型的概率密度函数是由这些Component线性组合而成,其公式如下:

(2.1)

{90% : 其中K为模型的个数,为第k个高斯的权重,为第k个高斯的概率密度函数,其均值为,方差为。 }{ 72% : 此概率密度的估计就是要计算、和各个变量。 }{100% : 当求出的表达式后,求和式的各项的结果就分别代表样本x属于各个类的概率。 }高斯混合模型的优点是投影后的样本点将获得每个类的概率,而非一个确切的分类标签。{ 56% : 相对于K-Means算法,高斯混合模型每一次的迭代计算量都比较大。 }由于高斯混合模型的聚类方法来源于EM算法,因此可能会产生局部极值,这与初始参数的选取密切相关。{ 64% : 高斯混合模型不仅可以用于聚类分析,同样也可用于概率密度估计。 }

基于密度的聚类算法

{ 77% : 基于密度的聚类分析算法的主要目是搜寻被低密度区域分割的高密度区域。 }{ 66% : 不同于基于

距离的聚类算法(基于距离的聚类算法是以数据集是球状簇为前提进行聚类的),基于密度的聚类算法可以发现任意形状的簇,这有利于处理带有噪声点的数据。 }在基于密度的聚类算法中,{ 64% : 分布在稀疏区域的对象通常被认为是噪声或边界点。 }

目前,{83% : 最流行的基于密度的聚类算法是DBSCAN(Density-Based Spatial Clustering of Application with Noise)[7]。 }在DBSCAN算法中数据点可以分为以下三类:

{ 73% : ①核心点:在-邻域内含有超过MinPts数目的点; }

②边界点:{80% : 在-邻域内点的数量小于MinPts,但是落在核心点的邻域内; }

③噪音点:{97% : 既不是核心点也不是边界点的点。 }

其中,为邻域半径,MinPts为指定的数目。{ 68% : DBSCAN的算法思想十分简单: }{83% : 若一个点p的-邻域包含多于MinPts个对象,那么创建以p作为核心对象的新簇; }{ 61% : 搜寻与核心对象直接密度可达的对象,将其合并;若没有新的点可以更新簇时,算法结束。 }

{ 62% : OPTICS是DBSCAN的变形[8], }它并不需要为范围参数选择合适的值,就产生与分层聚类相似的分层效果。Density-Link-Clustering结合单连通聚类和OPTICS的思想,完全消除了参数,并且通过使用R树索引增强了聚类性能。{ 77% : DBSCAN和OPTICS的主要缺点是它们是通过某种程度的密度下降来检测簇边界的。 } { 55% : 此外,它们无法检测现实生活数据中普遍存在的内在簇结构。 }而DBSCAN的变形方法EnDBSCAN可以解决此类问题[9]。Mean-shift方法基于核密度估计[10],将每个对象移动到其附近最密集的区域。最终,对象会收敛到密度的局部最大值。由于昂贵的迭代过程和密度估计,Mean-shift通常比DBSCAN的效率要低

表 2.1聚类分析算法对比表

Table 2.1 Clustering analysis algorithm comparison table

算法名称	参数	可伸缩型	用例	几何结构(度量标准)
K-Means	簇的个数	非常大的样本数,中等簇的个数	一般用途,簇的规模大致相等,平面几何,簇的数量不可过多	点对之间的距离
Affinity propagation	阻尼,样本首选项	在样本数方面不具有伸缩性	簇的数量较多,簇的规模有明显差异, 非平面几何	图距离
Mean-shift	带宽	在样本数方面不具有伸缩性	簇的数量较多,簇的规模有明显差异, 非平面几何	点对之间的距离
Spectral clustering	簇的个数	中等样本数,小型簇的个数	簇的数量不可过多,簇的规模大致相等,非平面几何	

图距离

Ward hierarchical clustering

簇的个数

大型样本数和簇的个数

簇的数量较多,可包含连通性约束条件

点对之间的距离

Agglomerative clustering

簇的个数,连通类型,距离

大型样本数和簇的个数

簇的数量较多,可包含连通性约束条件,非欧氏距离

任意点对之间的距离

DBSCAN

邻域大小

非常大的样本数,中等簇的个数

簇的规模有明显差异, 非平面几何

最近点之间的距离

Gaussian mixtures

参数较多

不具有可伸缩型

平面几何,有易于密度估计

到中心的马哈拉诺比斯距离

Birch

分枝, 阈值,可选的全局聚类器

非常大的样本数,中等簇的个数

大型数据集,离群点可移除,可数据简化

点对之间的欧式距离

近年来,诸多学者投身于提高现有算法性能的研究中[12][13],{ 63% : 其中包括CLARANS(Ng和Han,1994)[14][18]和BIRCH(Zhang等,1996)[15]。 }由于处理海量数据集的需求日益增长,所以研究人员试图通过换取聚类性能以增加所产生簇的语义分析能力,这一意愿引起了pre-clustering的发展,其中以canopy聚类最具代表性[16]。Canopy聚类算法可以处理超大型数据集,但是所得到的“聚类”仅仅是对数据集的粗略预分割,之后仍需使用现有的聚类分析算法对这些预分割数据集进行聚类。学者们一直在进行各种各样的聚类算法尝试,比如基于seed的聚类方法[17]。

聚类分析的评价标准

{ 58% : 每种聚类算法都有其各自的适用范围, }有的适用于小型数据集,{ 63% : 有的可以处理大型数据,有的可以发现任意形状的簇。 }但总的来说,数据挖掘针对聚类分析算法有以下评价标准:

①处理大型数据的能力

科学技术的发展使得获取的数据集的规模越来越大,那么能够处理大型数据集的能力已成为聚类分析算法的一般要求。目前,主流的聚类算法处理大型数据的主要方法通常是对大型数据进行随机取样获得一个较小型的数据集,然后再用聚类算法对样本数据集进行分析。{ 61% : 但是,这种方法有一个明显的缺点, }{ 63% : 样本数据集可能会导致不同的聚类结果, }难以对整体数据集进行真实客观的分析。{ 58% : 所以聚类算法的可伸缩性是现如今一个十分重要的研究内容。 }

②处理不同类型属性的能力

聚类分析算法不应该只能处理单一的数据类型,{ 62% : 而应该可以处理不同的数据类型,例如二元类型数据、

分类/标称类型数据、序数型数据等。 }

③可以发现任意形状的簇:

基于距离的聚类算法通常是通过区分对象之间的相似度进行聚类分析的,这类聚类算法趋向于发现数据规模相同或者类球形的簇。然而,现实生活中数据集分布通常是任意形状的。因此,{ 58% : 聚类算法应该具有对不同类型和不同密度的数据集的处理能力。 }

④初始化参数的数量最小化

聚类分析算法的参数选择一直是一个备受关注的问题,目前很多主流的聚类算法需要预先设置一些输入参数,例如生成簇的数量、近邻数、近邻半径等参数,才能对数据集进行聚类分析。通常来说,这些参数的选取在很大程度上影响着聚类结果的好坏。{ 63% : 如果聚类算法对参数选择十分敏感, }那么聚类结果的准确性将变得不稳定。因此,聚类算法应该减小参数的设置和参数对聚类结果的影响。

⑤处理异常数据的能力

异常数据通常是离群检测中的噪声数据或离群点,在离群检测中通常被视为是远离数据集中绝大多数数据对象的异常数据。现实生活中的数据集普遍存在着异常数据,数据集的噪声和冗余数据通常会使得聚类结果产生较大的差异。因此,能够处理异常数据变得十分重要。目前,一些算法对异常数据具有良好的处理能力,{ 57% : 例如DBSCAN聚类算法和基于密度的CURE聚类算法; }然而一些传统的聚类算法并不具备处理异常数据的能力,{ 55% : 例如ANGES聚类算法和K-Means聚类算法。 }

⑥处理高维数据的能力

{86% : 现在,随着数据集维数的不断增加, }聚类算法处理高维数据的能力也变得越来越重要。同时,由于高维数据的稀疏性和距离集中使得一些算法的聚类结果相较低维数据变得很差。虽然目前已提出一些处理高维数据的聚类算法,例如CLIQUE聚类算法,但是由于其聚类结果差异较大,所以在高维数据空间中CLIQUE聚类算法的应用并不广泛。{ 63% : 近年来,越来越多的研究人员开始关注高维数据, }{ 65% : 同时在高维数据空间中的聚类分析已成为目前聚类分析的一个主要方向。 }

⑦可解释性和可用性

{ 77% : 聚类分析是一种非监督式的机器学习方法, }对于要处理的数据集无法预先知道其统计分布情况,{ 79% : 所以当对数据集进行聚类分析后, }能够根据结果进行合理地解释,分析出数据的内在规律和模式。在低维数据空间中,例如2维或3维空间,可以通过绘图简单直观地展示聚类结果。然而在高维数据空间中,{ 64% : 如何解释聚类结果已成为目前聚类分析的一个主要方向。 }

聚类分析的评估检验

聚类结果的评估也被称为聚类验证。两个簇之间的相似性有诸多的检测方法。这些方法可以衡量不同的聚类方法对同一数据集的聚类效果。

内部检验

内部评估是指聚类结果的评估依赖于聚类的本身数据。{ 57% : 当聚类结果表现出高的类内相似性和低的类间相似性时, }这些评估方法会给出一个较高的分值。然而,具有高分值的内部评估却并不一定能够进行有效地信息检索[20]。另外,内部评估容易倾向于使用相同聚类模型的算法。比如,基于最优化对象间距离的K-Means算法,同样基于距离的内部评估将可能高估得到的聚类结果。因此,内部评估方法适用于比较两种算法的性能优劣,然而这却不包含有效性结果(valid results)的比较[19]。有效性指标依赖于数据集本身的结构。{ 59% : 比如,K-Means算法只能找到凸簇, }所以许多评估指标都是以此为假设的。同样,在具有非凸簇的数据集上,{ 58% : 既不使用K-Means 算法, }也不采用假定凸簇的评估标准。以下是基于内部标准的评估方法:

①Davies–Bouldin指数

Davies-Bouldin指数计算公式如下:

(2.2)

{ 56% : 其中,n为簇的个数,是簇的质心, }是簇中所有元素到簇质心的距离的平均值,是簇和簇之间的距离。因此,具有越低的Davies–Bouldin 指数则表明聚类的结果越好。

②Dunn指数

Dunn指数的目标是识别具有高密度且良好分割的簇群,{ 55% : 它是最小化簇间距离与最大化簇内距离的比值。 }{ 62% : Dunn指数的计算公式如下: }

(2.3)

{ 65% : 其中,表示簇i和簇j之间的距离, }为簇k的内距离。簇间距可以任意选择一种度量方式,比如,假定两个簇的中心之间的距离为两个簇之间的距离。同样,簇内距也有多重表示方式,比如,假定簇内的任意点对之间的最大距离为簇的簇内距。因此,具有越高的Dunn指数值则表明聚类的结果越好。

③轮廓系数

轮廓系数是簇内点对之间的平均距离与该簇内的点到其它簇的距离的最大值的比值[48],其计算公式如下所示:

(2.4)

其中,表示i向量到同一簇内其他点不相似程度的平均值,表示i向量到其他簇的平均不相似程度的最小值。可见轮廓系数的值总是介于[-1,1],越趋近于1代表内聚度和分离度都相对较优。{100% : 将所有点的轮廓系数求平均,就是该聚类结果总的轮廓系数。 }轮廓系数适用于K-Means算法,也可用于确定最优的聚类数。

外部检验

在外部检验中,{ 69% : 聚类结果的评估依赖于未进行聚类的数据, }例如已知的类标签和外部基准(external benchmark)。这些基准通常是由该方面的专家设置的一组预分类的元素。因此,这些基准集通常被视为是检验的黄金标准。这些检验方法用于比较聚类结果与预定基准类之间的近似程度。{ 62% : 由于类可能包含内部结构、属性不允许分离簇以及类可能包含异常情况等等, }这些因素导致研究人员对基准集是否能够对真实数据进行有效检验产生了疑问[21]。

这些方法与评估分类问题的方法相似。不同于统计被正确标记的类,这些方法统计的是同一个簇内点对之间相同标签的个数。以下是基于外部标准的评估方法:

①纯度

纯度用于衡量每个簇中包含单一类的个数[20],换言之,纯度是用于统计当前簇中最常见的类的样本点的个数。将所有簇的纯度累加并除以数据集的样本数就是该数据集的纯度。纯度的计算公式如下:

(2.5)

其中,M为簇集,D为类标签集,N为数据集的样本数。

②Rand指数

Rand用于衡量聚类簇与基准分类信息的相似度[22],也可视为该算法所作出的正确决策的百分比,其计算公式如下:

(2.6)

其中,{ 67% : TP为同一个类的点被分到同一个簇的数量,TN为不同类的点被分到不同簇的数量,FP为不同类的点被分到同一个簇的数量, }FN为同一类的点被分到不同簇的数量。Rand指数存在的一个问题是FP和FN具有相同的权重。{ 55% : 这对于某些聚类算法而言可能是不期望的特性, }下面的F-measure会解决这个问题。

③F-measure

F-measure通过参数来对召回度进行加权从而平衡FN的分布[23],{ 60% : 精度和召回度的计算公式如下: }

(2.7)

(2.8)

其中P是精度,R是召回度。结合精度和召回度,F-measure的计算公式如下:

(2.9)

其中,时,。换言之,当时,召回度对F-measure无影响。随着的增加,召回度在F-measure的权重也在增加。

本章小结

本章主要概述了聚类分析的相关信息,首先对聚类分析的定义作了详细说明,{ 57% : 接着介绍了几种主流的聚类模型, }主要包括连通性模型、中心性模型、分布模型、密度模型以及子空间模型等等;其次,{ 65% : 详细介绍了主流的聚类分析算法,主要包括层次聚类算法、基于中心的聚类算法、基于分布的聚类算法以及基于密度的聚类算法, }并且列出了代表算法的适用性和优缺点;再次,介绍了聚类分析常用的评价标准;最后,{ 64% : 详细介绍了聚类分析的评估检验方法, }主要包括内部检验和外部检验。

2 聚类分析概述

PCA-Hub 聚类算法

Hubness这一概念最早是由Milos˘ Radovanovic Ì•等人在2010年提出的[2],{ 66% : 现已被应用到机器学习、模

式识别等领域, }关于其应用最著名的是解决高维数据空间中的分类问题和聚类问题。Hubness分类问题的思想是给定一个数据集,构建k近邻矩阵,将矩阵中逆近邻数较多的样本点标记为hubs,然后再根据hubs的标签与它的近邻的标签的匹配程度分为good hubs和bad hubs,进而通过某种方式对它的近邻设置相应的权重,接着通过hubs的近邻更新hubs的值,最后对更新后的数据集进行分类处理。Hubness聚类问题的思想是首先构建k近邻矩阵,然后将矩阵中逆近邻数较多的点标记为hubs,接着在聚类迭代的过程中以某种方式将hubs设为当前簇的原型,最后对数据集进行聚类分析。虽然hub聚类算法可以解决传统聚类算法无法处理高维数据的问题,但是由于它未考虑高维数据空间中的冗余和噪声特征,从而降低了聚类性能。因此,本章通过探究逆近邻数的偏度与本征维度的相互关系,在原先的hub聚类算法之上以偏度的变化率为降维依据,保证了在对高维数据降维时不会损失过多的有价值信息,从而提高了聚类效果。本章将对hubness这一现象进行详细的阐述,并深入分析hub聚类算法,再此基础上提出了改进方法。

维数灾难

{ 60% : 维数灾难(Curse of Dimensionality),也称作维度的诅咒,这一术语最初是由Bellman在1961年考虑优化问题时引入的。 }维数灾难用于描述当数据空间的维数增加时,因其数据的体积呈指数型增长而遇到诸多问题的现象,并且该类现象不会出现在低维数据空间中。维数灾难在诸多领域引发了各种各样的问题,这些问题的共同之处在于随着数据维数的增加,{ 59% : 数据的体积将会呈指数型增长, }从而导致可用数据变得十分稀疏,而数据的稀疏性问题对于任何基于统计学的方法均是一个严峻的挑战。在机器学习领域的挑战中,通过从高维特征空间的有限训练数据中获得某种“自然状态(state of nature)”,{ 62% : 在训练样本的数量恒定时,随着维度的增加其预测能力逐渐减小,这通常称为Hughes影响[24]或者Hughes现象[25][26]。 }在距离度量领域的挑战中,高维数据空间的不同样本之间的距离变得基本相同。令高维欧几里德空间中超球体体积的计算公式为:

(3.1)

其中r为超球体的半径,{ 55% : d为数据集的维数;超立方体的计算公式为: }。当数据空间的维数趋向于正无穷时,超球体体积与超立方体体积的比值趋向于0,如等式3.2所示:

(3.2)

根据等式3.2可以看出,{ 56% : 从某种意义上而言在高维数据空间中几乎所有的数据都远离数据集的中心。 }同时,从另一个角度看通过比较高维数据空间中的最小值距离和最大值距离,可以看出当数据空间的维数趋向于无穷时,最小值距离和最大值距离趋向于相同,从而证明在高维数据空间中距离函数变得不再有意义,公式如下

(3.3)

在最近的研究中,Zimek等人归纳了在搜索高维空间数据时由于维数灾难可能会出现的问题:

①分数和距离的集中:用于区分数据样本的相关值(如距离)变得十分相似;

{ 57% : ②不相关的属性:在高维数据空间中大量的属性可能是不相关的; }

③参考集的定义:对于局部方法,参考集通常是基于最近邻的;

④无可比性的分数:不同的子空间会产生不具有可比性的分数;

⑤分数的可解释性:分数通常不再具有语义上的意义;

⑥指数搜索空间:搜索空间无法进行系统性地扫描;

⑦Hubness:某些对象容易频繁地出现在其它对象的近邻列表中。

目前,许多专门的方法只针对这些问题中的一个问题进行研究,留下很多开放性的问题值得我们继续分析讨论,在本章接下来的章节中将会对hubness这一现象进行深入研究。

Hubness 现象

在机器学习领域,受维数灾难影响的方法和任务包括贝叶斯建模(Bishop,2006)、最近邻预测(Hastie et al,2009)及搜索(Korn et al,2001)等。维数灾难造成的影响之一是距离集中(Distance Concentration),这是说在高维数据中点对之间的距离渐渐趋向于相同。Hinneburg和Aggarwal等人已经对高维数据中的距离集中和无意义的最近邻作了深入的研究。{ 57% : 维数灾难造成的另一方面影响是hubness。 }Hubness是一种某些对象容易频繁地出现在其它对象的k近邻列表中的现象。令表示一组数据点,其中为数据集D的元素。令dist表示在空间中的一个距离函数,其中如下定义:

(3.4)

在此基础之上,定义,表示为在空间中,样本点x出现在其它样本点的k近邻列表中的次数,也记为k-occurrence或

hubness score,仅仅根据数据点k-occurrence的大小无法确定hubness对实验结果有何种影响。数据点的bad k-occurrences表示为,是指样本点x作为数据集D中其它的样本点的k近邻次数,并且样本点x的标签和那些样本点的标签不匹配。数据点的good k-occurrences表示为,是指样本点x的标签与那些样本点的标签相匹配[2]。

k-occurrences的分布与维数的关系

为了研究分布与数据集维数的关系进行了以下的实验分析。通过使用k-occurrences分布的标准第三矩(也称作偏度)来表征的非对称性[27],

(3.5)

其中和分别是的均值和标准差。{ 65% : 偏度常常用于概率学和统计论中,衡量实数域中随机变量分布的不对称性。 }{ 59% : 偏度的值有正负之分,偏度为负则表明绝大多数的值(包括中值在内)位于平均值的右侧; }{ 62% : 偏度为正则表明绝大多数的值(不一定包括中值)位于平均值的左侧; }{ 64% : 偏度为零则表明数值近似地均匀分布在均值的两侧,却不一定为对称分布。 }偏度一般分为两种:

{ 64% : ①左偏态或负偏态:数据的主体集中在右侧, }左侧会呈现出较长的尾部;

{ 64% : ②右偏态或正偏态:数据的主体集中在左侧, }右侧会呈现出较长的尾部。

图 3.1 负偏态(左)和正偏态(右)

Figure 3.1 Negative skew(left) and positive skew(right)

若用三阶标准矩来表示随机变量X的偏度,那么偏度可被定义为:

(3.6)

其中为三阶中心矩,为标准方差,E为期望算子。等式最终以三阶累积量和二阶累积量的1.5次方的比值来表示偏度。偏度也可用非中心矩来表示,公式如下所示:

(3.7)

实验的数据集为从[0,1]均匀分布中随机抽取10000个维数为d的样本点,这些样本点之间彼此独立,采用以下三种距离度量方式来构建k近邻列表:欧几里德距离()、fractional 以及cosine[28]。图5(a-c)描述的是当时,的分布情况,其中数据集的维数d分别为:(a),(b),(c)。同样,图5(d-f)描述的是从正态分布中随机抽样获得的数据集的的分布情况。需要说明的一点是,只是一个经验值,当k取其它值时也可获得类似的结果。

图 3.2在不同维数、不同距离度量方式情况下样本数为10000的数据集的N5分布图

Figure 3.2 Distribution of N5 for Euclidean, l0.5, and cosine distances on data sets with n = 10000 points and dimensionality (a, d) d = 3, (b, e) d = 20, and (c, f) d = 100.

{ 57% : 从图3.2的描述中可以看出, }当数据集的维数时,在三种度量方式下的分布近似于二项分布(图5(a,d))。这表明在低维数据空间中,随机取样的点的双向图中度的分布近似于Erdo Ìs-Re Ìnyi(ER)随机图模型的度分布。然而随着数据集维数的增加,的分布将会逐渐偏离随机图模型的分布而且开始向右倾斜(图5(b,c)以及图5(e,f)其中度量方式为欧式距离和fractional)。通过上述实验可以观测到,在高维数据空间中的正偏态和hubs确实存在关联性,而且的偏度越大与之对应的数据集的hubness现象越强烈。

为了进一步研究的偏度与数据集维数的相关性,{ 66% : 采用斯皮尔曼等级相关系数进行评测。 }斯皮尔曼等级相关系数(Spearman correlation)是用于评估两个变量相关性的非参数指标[29],记作。对于样本数为n的数据集,为其对应的等级数据,那么相关系数的公式如下:

(3.8)

由于在现实应用中变量之间的连结并没有显著作用,因此可以对进行如下简化[30]:

(3.9)

其中,表示被评估的两个变量等级之间的差值,n为样本数。斯皮尔曼相关系数阐述了X(独立变量)与Y(依赖变量)的相关性。若变量X增加时,{ 56% : 变量Y也增加,那么斯皮尔曼相关系数的值为正数;若变量X增加时,变量Y却在减少,那么斯皮尔曼相关系数的值为负数; }若变量X和变量Y没有相关性,{ 64% : 那么斯皮尔曼相关系数则为零。 }Milosˇ Radovanovic Ì•等人研究了50多个真实数据库的维数与其的斯皮尔曼相关系数(0.62),这一结果进一步说明数据集的维数和hubness现象存在着强烈的正相关性。

Hubs的位置

为了进一步探究hubness这一现象,需要对hubs位置的分布进行深入分析。数据集仍然采用之前的均匀分布和正态分布随机取样数据。以样本数据分布的均值作为参考点,可以观测到样本点的k-occurrences值与其在数

据空间中分布的相互关系。图6描述了在不同维数下(,,)每个数据样本x的值与其距离样本均值的相关性,其中。{ 65% : 从下图中可以看出,随着数据集维数的增加, }与该样本到数据集均值的距离出现了强烈的负相关性,这意味着越接近样本均值的点越有可能为hubs。{ 57% : 值得注意的是,只是一个经验值, }当k为其它值时也可获得相似的结果。该实验结果表明在高维数据空间中,当潜在的数据分布是单峰时,hubs会接近整体样本分布的均值;而当潜在的数据分布为多峰时(如多个单峰分布混合而成),hubs趋向于最近的单峰分布的均值。

图 3.3在不同维数、不同距离度量方式情况下样本数为10000的数据集的N5(x)分布图

Figure 3.3 Scatter plots and Spearman correlation of N5(x) against the Euclidean distance of point x to the sample data-set mean for data sets with (a, d) d = 3, (b, e) d = 20, and (c, f) d = 100.

到目前为止,hubness现象对机器学习应用的影响还没有进行彻底的研究。在接下来的章节中,将会针对聚类分析这一领域对hubness进行深入的研究。

Hub聚类算法分析

基于距离的聚类算法的主要目标是最小化同一个簇内对象之间的距离同时最大化簇间对象之间的距离。然而在高维数据空间中,无论是在efficiency还是effectiveness上,高维数据都对传统聚类算法造成了实质上的困难。因此,必须使用一种新的技术来对高维数据进行聚类分析。通常的想法是将原始的高维数据映射为一个较低维的流型结构[31],然后再进行聚类分析,该思想的代表算法是子空间聚类算法。在许多实际应用中,例如文本聚类和主题检测[32][33],为了进行有意义地聚类分析通常会将原始数据投影到某些较低维的子空间和流型结构(manifolds)中。{ 59% : 一般来说有两种类型的子空间聚类方法, }{ 59% : 一种方法是试图找到一个真实的特征子空间, }另一种方法是在模拟过程中自动对特征进行加权处理以增加聚类效果。从理论上来说,在较低维的划分子空间中执行标准聚类算法看似是可行的[34]。然而,如果划分出的较低维子空间并不是真正意义上的低维数据,那么标准的聚类算法是无法处理的。此外,许多子空间聚类算法是基于密度的聚类或者是K-Means算法的扩展,而基于密度的聚类算法和K-Means聚类算法均不适用于高维数据聚类分析。由此可见,子空间聚类方法在高维数据空间中也存在诸多限制。因此,在接下来的章节中将会采用一种不同的方法,通过利用最近提出的k近邻图中出现的hubness现象进行聚类分析。

在高维数据空间中,hubness现象将会对基于距离的聚类算法造成两方面影响。一方面,具有低k-occurrences的样本点很可能会增加簇内对象之间的距离,使得这些点远离数据集的其它点,可以将其视为离群点。目前,关于离群点在聚类分析方面的应用已经作了诸多的研究,通常离群点被发现之后会直接将其移除。另一方面,具有高k-occurrences的样本点,也就是hubs,很有可能会接近簇的中心。值得注意的是,一些聚类算法因为hubs的存在而使聚类性能变差,这是因为某些hubs会接近来自不同簇的样本点[2]。之前已经提到过,相比其它样本点而言,k-occurrences值越高的样本点越容易接近簇的均值,随之而来便产生了一个疑问:hubs会是当前簇的中值样本吗?Nenad Toma sev等人通过实验研究发现[35]:在低维数据空间中,hubs远离簇的中心甚至远离普通的点。{ 75% : 然而,随着数据集维数的增加, }簇的中心到hubs的最小距离会逐渐收敛于簇的中心到簇的中值样本的最小距离。这表明一些簇中值样本就是hubs。然而,簇的中心到hubs的最大距离却没有上述的相关性。同时观测到随着每一次的聚类迭代,簇的中心到hubs的最大聚类也逐渐减小,这就表明簇的中心越来越接近hubs。因此在高维数据中,hubs可以在很大程度上代表该簇中的元素。

图 3.4 通过一个简单的例子说明在划分聚类迭代过程hubs原型与簇中心原型或簇中值原型的差别:红色虚线的圆圈代表簇的中心(C),黄色的点状圆圈代表簇的中值样本(M),绿色圆圈代表两个hubs(H1,H2),其中最近邻居数为3。

Figure 3.4 An example of the difference between using hubs and centroids/medoids as cluster prototypes in partitional clustering iterations: The red dashed circle marks the centroid (C), yellow dotted circle the medoid (M), and green circles denote two elements of highest hubness (H1,H2), for neighborhood size 3.

既然hubs可以被视为一种度量局部中心性的方法,那么可以采取多种方式将hubs应用到各种聚类分析中。{ 60% : 在K-Means迭代过程中, }簇中心依赖当前簇中的所有元素,而hubs仅依赖它们的近邻元素,因此hubs携带着很多局部的中心性信息。Hubs主要可分为全局hubs和局部hubs。局部hubs是全局hubs在给定任一簇情况下的约束。因此,局部hubs的k-occurrences值是指在同一个簇中的某个样本点的k-occurrences的数量。同时,簇的中心和簇的中值样本容易趋向于k-occurrences值较大的样本点,也就是hubs,这意味着使用hubs作为簇原型可以加快算法的收敛速度。为了解释这一点我们设计了一个简单的模型如图3.4所示,图3.4通过2维数据空间模拟了高维数据空间中经常出现的hubness现象,该图阐释了以hubs作为簇原型不仅可以加快算法的收敛速度而且有助于发现更好的簇结构。Tomasev等人提出了基于hubness的K-Means扩展聚类算法[35]。

在global K-hubs(GKH)算法中,hubs取代簇中心作为每次迭代过程中的簇原型。初步实验表明GKH方法容易

过早地收敛到次优的簇结构。因此,{ 56% : 在GKH算法中引入了随机因子, }在每次迭代过程中hubs与其它样本点以某种概率被选为簇原型,该概率依赖样本点本身的值。这种算法被称为global hubness-proportional clustering(GHPC)。同样值得注意的是,相比基于密度的聚类分析,将hubs作为簇原型进行聚类分析的优点不仅仅是hubs能够很好地反映局部簇的中心性,而且hubs对数据集的规模不敏感。当数据集是由若干个密度差异较大的簇组成时,这一性质变得至关重要。向上或向下伸缩簇并不会改变其近邻结构,通常来说,也不会改变数据集的hubness特性。自然地,将hubs作为簇原型进行聚类分析并不一定能够对于任何数据集都可以得到最优的簇结构。有时,将簇中心作为簇原型反而可以得到不错的簇结构。{ 57% : 因此,需要对GHPC聚类算法进行改进和扩展: }在确定性迭代过程中使用簇中心作为簇原型;在随机性迭代过程中使用使用样本的随机概率作为簇原型。这种混合的方法叫做global hubness-proportional K-means(GHPKM)。由于GKH、GHPC和GHPKM聚类算法只能够发现超球形的簇[36],所以将核方法引入hubs聚类算法以便可以发现不同类型的簇结构。{ 56% : 下面将会详细介绍不同的hub聚类算法。 }

Deterministic方法

使用hubs进行聚类分析的一种简单方法是将hubs作为每次迭代过程中当前簇的簇原型,该算法一般称为K-hubs算法,其算法思想如下:

Algorithms 1. K-hubs

initializeClusterCenters();

Cluster[] clusters = formClusters();

repeat

â€for all Cluster c clusters do

â€DataPoint h = findClusterHub(c);

â€setClusterCenter(c, h);

â€end for

clusters = formClusters();

until noReassignments

return clusters

{ 62% : 尽管K-hubs聚类算法可以得到很好的聚类效果, }{ 57% : 但是它对初始簇原型十分敏感, }而且容易获得次优的聚类结构。为了增加找到全局最优解的概率,下面将介绍随机变量的K-hubs聚类分析算法。

Probabilistic方法

尽管拥有最高k-occurrences值的样本点可以最大可能地代表当前簇的信息,但是簇中其它样本点也有可能包含该簇的重要信息。因此,在K-hubs算法上对簇原型的选择加入了一定的随机性,通过使用模拟退火方法实现了一个平方hubness-proportional的随机方法[37],将温度因子引入到K-hubs算法中,那么它的初始簇原型就是完全随机的,该方法称为hubness-proportional clustering(HPC)聚类算法,其算法思想如下:

Algorithm 2. HPC.

initializeClusterCenters();

Cluster[] clusters = formClusters();

float t = t0; initialize temperature

repeat

â€float = getProbFromSchedule(t);

â€for all Cluster c clusters do

â€if randomFloat(0,1) ^ then

â€DataPoint h = findClusterHub(c);

â€setClusterCenter(c, h);

â€else

â€for all DataPoint x c do

â€setChoosingProbability(x,);

â€end for

â€normalizeProbabilities();

â€DataPoint h = chooseHubProbabilistically(c);

â€setClusterCenter(c, h);

â€end if

â€end for

â€clusters = formClusters();

â€t = updateTemperature(t);

until noReassignments

return clusters

在高维数据空间中,Hubness-proportional 聚类分析算法的可行性在于k-occurrences偏度的统计分布。在k-occurrences偏度的统计分布中,绝大多数的点拥有较低的值,这通常意味会它们会被GHPC算法忽略掉,因为它们被视为是十分差的簇原型的候选者而且它们被选择的概率也非常低。关于选择样本点的方法,GHPC算法采用了一个相当繁琐的温度因子方案,当然其它的随机方案也是可行的,甚至会产生更好的聚类效果。

Hybird方法

K-hubs聚类算法和GHPC聚类算法都没有关注数据或对象的表现形式(representation),它们只在意样本间的距离矩阵。然而,{ 64% : 如果数据的表现形式是已知的, }那么便可以利用簇中心的相关性质进行聚类,同时使用样本点的k-occurrences指导聚类搜索,最终会形成一个基于簇中心的聚类结构。该算法被称为hubness-proportional K-means(HPKM)聚类算法,它与GHPC聚类算法的唯一不同之处在于迭代过程中的确定阶段使用的是K-Means更新簇原型而非K-hubs。

Algorithm 3. HPKM.

initializeClusterCenters();

Cluster[] clusters = formClusters();

float t = t0; initialize temperature

repeat

â€float = getProbFromSchedule(t);

â€for all Cluster c clusters do

â€if randomFloat(0,1) ^ then

â€DataPoint h = findClusterCentroid(c);

â€setClusterCenter(c, h);

â€else

â€for all DataPoint x c do

â€setChoosingProbability(x,);

â€end for

â€normalizeProbabilities();

â€DataPoint h = chooseHubProbabilistically(c);

â€setClusterCenter(c, h);

â€end if

â€end for

â€clusters = formClusters();

â€t = updateTemperature(t);

until noReassignments

return clusters

Kernel GHPKM方法

K-hubs、GHPC和GHPKM算法的主要缺陷是它们只能发现发现超球面的簇。{ 60% : 然而在现实生活中簇的形状是任意的, }所以需要寻找一种新的技术来解决此问题。Kernel K-Means算法[38]是K-Means算法的一个扩展,同样K-hubs、GHPC和GHPKM算法也是K-means的扩展,因此可以将这些算法与kernel方法结合起来进行聚类分析。令表示数据集的簇,其中,使用非线性函数,那么kernel GHPKM聚类算法的目标函数如下:

(3.10)

{ 62% : 其中,为每个样本点所对应的权重, }为每个簇的原型,这是kernel K-means算法和kernel GHPKM算法的第一个不同之处:{ 71% : 在kernel K-means算法中, }并不一定是簇的中心,也可以是hubs或者其它的样本点。然而,{ 59% : 随着迭代次数的不断增加,这个不同之处将会越来越小甚至消失。 }这一现象是由模拟退火算法的降温方法所引起的,{ 66% : 随着迭代次数的增加随机性发生的概率将越来越小, }最终演变成确定性的迭代过程。因此,Kernel GHPKM算法可以使用与kernel K-Means算法相同的最小化函数,同时也可以通过随机选择初始值来避免局部最优值的问题。令为簇c的中心,在映射函数下,簇中心可通过下面的公式(3.11)计算得出。有时为了区分簇原型和簇hub原型,我们通常用和分别标记。

(3.11)

从下面的等式中可以看出,簇中心的最优化目标是通过最小化加权映射距离的平方和获得的。虽然使用hubs作为簇原型并不会最小化平方距离和,但是它却会带来其它的益处。

(3.12)

{ 60% : Algorithm 4. Kernel GHPKM. }

initializeClusterCenters();

float t = t0; initialize temperature

repeat

â€¢loat = getProbFromSchedule(t);

â€¢or all Point x dataset do

â€¢closestCluster = NULL;

â€¢minimalDistance = MAX VALUE;

â€¢or all Cluster c clusters do

â€¢f getClusterCenter(c) NOT NULL then

â€¢distance = getDistanceToHub(c);

â€¢f distance minimalDistance then

â€¢minimalDistance = distance;

â€¢closestCluster = c;

â€¢nd if

â€¢lse

â€¢distance = getDistanceToCentroid(c);

â€¢f distance ≤ minimalDistance then

â€¢minimalDistance = distance;

â€¢closestCluster = c;

â€¢nd if

â€¢nd if

â€¢nd for

â€¢assignPointToFutureCluster(x, closestCluster)

â€end for

â€updateClusterAssignments();

â€for all Cluster c ∈ clusters do

â€if randomFloat(0,1) ^ then

â€setClusterCenter(c, NULL);

â€else

â€for all DataPoint x c do

â€setChoosingProbability(x,);

â€end for

â€normalizeProbabilities();

â€DataPoint h = chooseHubProbabilistically(c);

â€setClusterCenter(c, h);

â€end if

â€end for

â€ = updateTemperature(t);

â€calculateErrorFunction();

until convergenceCriterion

clusters = formClusters();

return clusters

本节详细地介绍四种hub聚类算法,{ 61% : 并归纳出了各自的优缺点和适用范围。 }从上述的阐述中可以发现,以特定方式将hubs最为簇原型不仅可以获得更好的簇结构同时也可以加快聚类分析的收敛速度。从本质上说,如果在聚类分析算法的迭代过程中当前的簇是由多个紧凑的部分组成,那么簇的中心和簇的中值样本并不一定能够代表有意义的原型。理想情况下,在搜索最佳的划分和最优簇结构时,我们希望在每一次的迭代过程中都能将独立的子部分分给不同的簇。然而,以簇中心和簇中值样本作为簇原型很可能会减小迭代之间的差异,增加迭代次数甚至得到次优的簇结构。此外,多组分簇(multi-component clusters)的簇中心和簇的中值样本并不能与局部组分簇中心对应。因此,使用hubs作为搜索原型可以克服在高维数据空间中进行聚类分析的相关问题。虽然hub聚类算法相比经典聚类算法中在高维数据空间中表现出了显著优势,然而它却没有关注高维数据空间中的冗余和噪声数据,{ 63% : 因此并未获得更优的聚类效果。 }{ 55% : 下面的章节将会针对这一问题进行深入的分析研究, }并提出可行的更改方案。

Hub聚类算法的改进

基于距离的聚类算法的主要目标是最小化同一个簇内对象之间的距离同时最大化簇间对象之间的距离。在高维数据空间中,样本的k-occurrences偏度将会对上述两个对象造成影响。一方面,具有低k-occurrences的样本点很可能会增加簇内对象之间的距离,这些样本点远离数据集的其它点,可以将其视为离群点。目前,关于离群点在聚类分析方面的应用已经作了诸多的研究,通常离群点被发现之后会直接将其移除。另一方面,具有高k-occurrences的样本点,也就是hubs,很有可能会接近簇的中心。另外,数据集的hubness度依赖于数据集的本征维数而非嵌入维数(embedding dimensionality)[39][40]。本征维数(Intrinsic dimensionality)是指表示数据集所有点对之间的距离所需特征的最小数量[41]。通常,hubness与本征维数相关而与距离或相似度的度量方式无关。通常,较低的k-occurrences值表明该样本点远离数据样本中的其它点,并且很有可能是一个离群点。然而,在高维数据空间中,由于数据本身的分布情况使得较低的k-occurrences样本点变得很普遍,这些样本点将会增加簇内样本之间的距离。同样值得注意的是,一些聚类算法因为hubs的存在而使聚类性能变差。这是因为某些hubs会接近来自不同簇的点[35]。之前已经提到过,相比其它点而言,k-occurrences值越高的样本点越容易接近簇的中心,Nenad Toma sev等人通过实验研究发现在高维数据空间中,hubs可以在很大程度上代表当前簇中的元素,从而获得更好的聚类结构和更快的聚类收敛速度。

虽然hub聚类算法利用了hubs在高维数据空间中的特性,并获得了较为不错的聚类效果,然而它却没有考虑高维数据空间中的冗余和噪声数据,{ 63% : 因此并未获得更优的聚类效果。 }接下来的章节将会通过分析的偏度与本征维数的相互关系,探究降维技术是否能够缓解的偏度等问题,{ 60% : 从而进一步对hub聚类算法进行改进, }其中降维技术使用的是主成分分析方法。

主成分分析

在多变量的统计分析中,{ 73% : 主成分分析(Principal components analysis,PCA)常常用于分析和简化数据集[42]。 }主成分分析通过保留对方差贡献最大的样本特征,从而降低数据集的维数。Pearson于1901年发明了主成分分析[43],常常用于数据分析以及模型建立。主成分分析的主要思想是将协方差矩阵进行特征分解,从而获得数据的主要成分(特征向量)及其对应的权重(特征值)。{ 56% : 使用主成分分析算法对数据集进行降维处理后可以极大地提升无监督特征学习的速度。 }以下是主成分分析算法的具体思想:

①使用n行d列的矩阵X表示原始数据;

②将矩阵X的每一列进行零均值化,即减去这一行的均值;

③求解协方差矩阵;

{ 70% : ④求解协方差矩阵的特征值及其特征向量; }

⑤令特征向量按照其对应的特征值降序排序,取前k列组成新的矩阵P;

⑥即为降维后新的数据。

主成分分析基于最大方差矩阵理论,{ 56% : 通过协方差矩阵的特征向量选择k维理想特征, }也就是说,{ 72% : 在减少数据集维数的同时保留数据集中对方差贡献最大的特征。 }{100% : 这是通过保留低阶主成分,忽略高阶主成分做到的。 }{100% : 这样低阶成分往往能够保留住数据的最重要方面。 }{80% : 主成分分析主要是通过对协方差矩阵进行特征分解,以得出数据的主成分(即特征向量)与它们的权值(即特征值)。 }{93% : 这可以理解为对原数据中的方差做出解释:哪一个方向上的数据值对方差的影响最大? }{ 72% : 换而言之,PCA提供了一种降低数据维度的有效办法; }如果分析者在原数据中除掉最小的特征值所对应的成分,那么所得的低维度数据必定是最优化的(也即,这样降低维度必定是失去讯息最少的方法)。

{ 73% : 如何选择k值,即保留多少个主成分? }对于k值的选择,通常以k值所保留的方差百分比作为参考依据。一般来说,{ 64% : 当时保留了百分之百的方差,也就是说原先数据的所有变化均被保留了下来;相反, }当时只保留了百分之零的方差。{ 61% : 通常而言,令表示协方差矩阵的特征值(从大到小排列),特征值对应的特征向量为, }{ 65% : 如果选择k个主成分那么保留的方差百分比可表示为: }

(3.13)

通常而言,通过选择最小的k值使得保留方差的范围位于90~98%之间,在不同的应用领域中这个范围可自行调整。

基于偏度的降维方法

关于数据降维的方法有多种,本文采用的是主成分分析法。{91% : 主成分分析经常用于减少数据集的维数,同时保持数据集中的对方差贡献最大的特征。 }当没有任何假设信息的信号模型时,主成分分析在降维的同时并不能保证信息的不丢失,其中信息是由香农熵来衡量的。然而,香农熵却无法作为数据有效降维时的衡量标准,因此本文采用了的偏度这一指标。下文中将会探讨在使用降维技术PCA的情况下的偏度和本征维数的相互作用。此研究的主要目的在于探讨降维是否能够缓解的偏度这一问题。 “因为观察到的偏度与本征维数强烈正相关,本征维数对到数据集的均值或到最接近簇的均值有着积极影响,这意味着在较高(本征)维数的数据集中,hubs变得越来越接近数据集的中心或者最接近的簇的中心” 。{ 57% : 实验过程中采用的距离度量方法是闵可夫斯基距离(Minkowski distance)[44],它是一种非常常见的衡量样本点之间距离的方法, }假设数值点P和Q坐标如下:

{ 72% : 那么,闵可夫斯基距离定义为: }

(3.14)

该距离最常用的p值是2和1,{ 58% : 前者是欧几里得距离(Euclidean distance)[45], }{ 73% : 后者是曼哈顿距离(Manhattan distance)[46]。 }可夫斯基距离虽然比较直观明了,但是它并没有考虑数据的统计分布,因此具有某些局限性。例如,若x方向上的幅值比

y方向的幅值要大得多,那么闵可夫斯基距离将会在很大程度上扩大x方向上的作用。因此,在计算对象之间的距离之前,根据数据的分布情况可能需要进行z-transform处理,即减去该维度上的均值,并除以其标准差:

(3.15)

其中,是当前维度上的均值,是当前维度上的标准差。由此可见,z-transform是基于数据在各维度上不相关的假设,并利用数据的统计分布特性进行不同的距离度量的。

为了探究在使用降维技术的情况下的偏度和本征维数的相互作用,本文使用了来自加州大学尔湾分校(UCI)机器学习库[47]的数据集进行观测的分布。在表3.1中包含了以下信息:数据集的名称(第1列);数据集的样本数(n,

第2列);数据样本的特征维数(d,第3列);数据集簇的个数(cls,第4列)。

表 3.1来自UCI机器学习库的真实数据集

Table 3.1 Real data sets from UCI Machine Learning Repository

数据集
样本数
维数
簇的个数
arrhythmia
452
279
10
Ionosphere
351
34
2
mfeat-factors
2000
216
10
mfeat-fou
2000
76
10
musk
476
166
2
spectrometer
531
100
10
sonar
208
60
2

图3.5描述了针对若干个真实数据集(musk,sonar,mfeat-fou等)通过降维方法获得的维数占原有数据集维数的百分比与之间的相互关系。数据之间距离的度量方法为Minkowski距离,其中p的取值分别为:2(欧几里得距离)。从左往右观察,对于大部分数据集而言利用PCA降维算法,保持相对恒定直到降维后留下特征的百分比较小时才会陡然下降。因此,当达到数据集的本征维数时若继续减小维数则会导致有价值的信息丢失。{ 64% : 针对PCA方法对数据进行降维时, }若降维后本征维数未发生明显变化,那么降维并不会对hubness这一现象有显著影响。

图 3.5 N10的偏度与降维维数的关系

Figure 3.5 Skewness of N10 in relation to the percentage of the original number of features maintained by dimensionality reduction.

PCA-Hub聚类算法

上节通过实验研究发现,数据集的偏度与数据集的维数存在强烈正相关,更确切地说,是与数据集的本征维数存在强烈正相关,而数据集的本征维数表示的是数据集所有点对之间的距离所需特征的最小数量[41]。因此,{ 58%：本文提出了PCA-Hub聚类算法, }此算法以数据集的偏度作为降维衡量标准,通过不断减小数据集的原始维数来逐渐逼近数据集的本征维数,这样不仅不会损失数据集的“原始”信息,而且还能消除其中的冗余和噪声数据,有利于更快的发现更优的簇结构。下面是PCA-Hub聚类算法的具体步骤:

①数据预处理

实验之前首先要观察数据并获知数据的特性,并且应该针对具体的数据采取合适的预处理技术。本章采用的数据预处理技术为一种常见的数据归一化方法------逐样本均值消减(也被称为移除直流分量,局部均值消减,消减归一化),即对于每个样本点减去数据统计分布的平均值[53]。

②构造KNN邻域矩阵

基于距离的聚类算法需要考虑不同的距离度量方法对于聚类性能的影响,不同类型的数据集应该采用各自适合的距离度量方法。在确定合适的距离度量方式之后,需要选定合适的近邻数k用于构建KNN邻域矩阵。

③计算逆近邻偏度

通过KNN邻域矩阵可获得每个样本点的逆近邻数,通过偏度可以衡量样本逆近邻数的非对称性,并以此分析数据集的hubness情况。

④PCA降维

因为逆近邻的偏度与数据集的本征维数强烈正相关,因此将偏度作为主成分分析的降维指标,当偏度小于某一设定的阈值时便可认为数据集已损失了较多的本征维数,即剩下的维数为理想的k个主成分。

⑤聚类分析

由于hub可以代表局部中心性,{ 61%：所以可以用hub聚类算法对降维后的数据进行聚类分析。 }

图 3.6 PCA-Hub聚类算法流程图

Figure 3.6 PCA-Hub clustering algorithm flow chart.

实验结果及其分析

本小节将会从三个方面分别探讨PCA-Hub聚类算法的聚类性能,具体情况如下:

①PCA-Hub聚类算法的轮廓系数

实验数据来源于加州大学尔湾分校(UCI)机器学习库。表3.2中第5列为真实数据集的偏度值,其中10代表k近邻数。从表中数据可以看出,对于大多数数据集的的分布发生了倾斜。虽然k的值是固定的,但是使用其它的k值也可得到类似的结果。采用轮廓系数作为聚类结果的评测指标[48]。本文方法与KMEANS[35]、GHPKM[35]、Ker-KM[36]和 Ker-KM[36]方法进行了比较,其中PH-KM为本文的聚类方法。实验结果如表3.2所示,下表中加粗的数据表示当前数据集的最优值。

表 3.2 UCI库中数据集的聚类质量(轮廓系数)

Table 3.2 Clustering quality expressed as silhouette index on data sets from the UCI repository.

数据集

样本数

维数

簇的个数

距离度量

KME

ANS

[9]

GHPKM

[9]
Ker-KM [4]
Ker-GHPKM
[4]
PH-KM
Ionosphere
351
34
2
1.72
l2
0.28
0.28
0.28
0.25
0.41
mfeat-factors
2000
216
10
0.83
l2
0.18
0.20
0.17
0.18
0.24
musk
2000
166
2
1.33
l2
0.28
0.28
0.29
0.29
0.31
parkinsons
195

22
2
0.73
l2
0.42
0.44
0.64
0.21
0.88
sonar
208
60
2
1.35
l2
0.20
0.21
0.26
0.17
0.22
wpbc
198
33
2
0.86
l2
0.16
0.16
0.32
0.22
0.31
AVG-UCI
0.25
0.26
0.33
0.22
0.39

图 3.6 UCI库中数据集的聚类质量(轮廓系数)

Figure 3.6 Clustering quality expressed as silhouette index on data sets from the UCI repository.

对于每一个数据集而言,取KMEANS、GHPKM、Ker-KM以及Ker-GHPKM聚类算法中轮廓系数的最大值作为经典聚类算法的最优值,然后同本文的PH-KM聚类算法进行比较。{ 73% : 实验结果表明,相比之前的聚类

算法, }本文提出的PH-KM聚类算法在轮廓系数上平均提高了15%。{ 69% : 从表3.2的实验结果可以看出, }经典的KMEANS聚类算法更适用于低维数据聚类;在数据集缺乏hubness特性的情况下,GHPKM、Ker-GHPKM等hub聚类算法表现不佳,其性能接近于KMEANS算法;然而当数据集呈现出较高的hubness特性时,GHPKM、Ker-GHPKM等hub 聚类算法的表现要优于KMEANS算法。同时,本文提出的 PCA-Hub聚类算法无论数据集是否呈现出较高的hubness特性,均可以取得不错的聚类效果,相比之前的聚类算法适用范围更广,聚类性能更佳。

②PCA-Hub聚类算法对近邻数k的敏感程度

图 3.7 PCA-Hub聚类算法对近邻数k的敏感程度

Figure 3.7 The sensitivity of PCA-Hub Clustering Algorithm to k nearest Neighbors.

由于PCA-Hub聚类算法是基于K-Means聚类算法在高维数据空间的扩展方法,因此有必要研究其对于近邻数的敏感程度。实验所用的数据集和距离度量方法仍然保持不变,PCA-Hub聚类在每个数据集上均重复聚类50次,近邻数k的取值范围从5到25。图3.7为实验结果示意图,从图中可以看出当数据集的维数较低且的偏度也不高时,PCA-Hub聚类算法对近邻数k这一参数的选择表现出了明显的依赖性,{ 59% : 聚类算法的性能在很大程度上取决于近邻数的取值;同时, }图3.7表明当数据集本身的维数较高时或者的偏度不低时,PCA-Hub聚类算法在使用不同的近邻数k时表现出了相似的聚类性能,因此近邻数的选择对于PCA-Hub聚类算法的聚类结果影响并不强烈。

③PCA-Hub聚类算法聚类结果的一致性

为了研究PCA-Hub聚类算法结果的稳定性或一致性,本文进行了如下的实验研究:采用之前的UCI数据库,设置参数近邻数k为最优值,聚类算法的重复次数为50次,并记录每一次的聚类结果。从图3.8中可以看出,在实验环境和聚类分析算法参数一致的情况下,PCA-Hub聚类算法在开始的一小段重复次数时发生了些许的波动,但随着聚类算法重复次数的增加,聚类结果渐渐趋于稳定,并最后收敛于某一个恒定的值,这一现象表明PCA-Hub聚类算法的聚类性能,尤其是聚类重复次数比较高的情况下,在很大程度上具有一致性。

图 3.8 PCA-Hub聚类算法对近邻数k的敏感程度

Figure 3.8 The sensitivity of PCA-Hub Clustering Algorithm to k nearest Neighbors.

本章小结

本章节首先对高维数据空间中的维数灾难做了简要的分析介绍,并归纳出了维数灾难在机器学习中的影响。针对这种现象引入了hubness这一较新的概念,并对hubness做了十分详尽的描述:首先给出了hubness现象相关的定义,并分析了hubs在数据集中的位置;然后根据hubs的中心性特征介绍了hub聚类算法,并归纳出了其优缺点。通过研究发现虽然hub聚类算法可以在高维数据空间中进行聚类分析,但是却忽略了高维数据中的冗余和噪声数据,从而导致聚类效果不佳。随后,通过的偏度来表征数据集的hubness特性,并以的偏度与本征维数强烈正相关为理论基础,通过构建数据集的KNN邻域矩阵,以偏度的变化率作为降维依据选出理想的k个主成分,之后再对降维后的数据集进行聚类分析。最后通过多次实验分别从聚类结果的好坏(轮廓系数)、对近邻数k的敏感程度和聚类结果的一致性三方面进行了深入分析,{ 61% : 实验结果表明,在聚类结果方面, }无论数据集是否呈现出较高的hubness特性,本章提出的PCA-Hub聚类算法均可以取得不错的聚类效果,相比之前的聚类算法,轮廓系数平均提高了15%;在对近邻数k的敏感程度方面,PCA-Hub聚类算法在数据集本身的维数较高或者的偏度不低时,对近邻数k的选择表现不强烈;在聚类结果的一致性方面,PCA-Hub聚类算法在实验环境和聚类算法参数一致的情况下,聚类结果在很大程度上具有一致性。

错误!未找到引用源。错误!未找到引用源。PCA-Hub 聚类算法

3 PCA-Hub聚类算法

Quick PCA-Hub聚类算法

在第三章PCA-Hub聚类算法分析中,以偏度的变化率作为降维依据,利用主成分分析降维方法对数据集进行降维的同时尽可能地保留了数据集的本征维数,从而提高了聚类算法的性能。虽然PCA-Hub聚类算法可以解决高维数据中的冗余和噪声特征,并且降维后的数据集也可以加快聚类分析的速度和获得不错的簇结构,但是在获取主成分分析方法的k值时,尤其对高维数据而言,该阶段的计算代价过于昂贵。因此需要找到一种可以快速获得主成分分析方法中理想k值的算法,本章将会从快速搜索k个主成分和快速搜索最近邻居两方面介绍加快PCA-Hub聚类算法的速度,其中图4.1为Quick PCA-Hub聚类算法的流程图。

图 4.1 Quick PCA-Hub聚类算法流程图

Figure 4.1 Quick PCA-Hub clustering algorithm flow chart.

快速搜索k个主成分

Quick PCA-Hubness聚类算法的整体流程如下所示:

①数据预处理

实验之前首先要观察数据并获知数据的特性,并且应该针对具体的数据采取合适的预处理技术。本章采用的数据预处理技术为一种常见的数据归一化方法------逐样本均值消减(也被称为移除直流分量,局部均值消减,消减归一化),即对于每个样本点减去数据统计分布的平均值[53]。

②构造KNN邻域矩阵

基于距离的聚类算法需要考虑不同的距离度量方法对于聚类性能的影响,不同类型的数据集应该采用各自适合的距离度量方法。在确定合适的距离度量方式之后,需要选定合适的近邻数k用于构建KNN邻域矩阵。

③计算逆近邻偏度

通过KNN邻域矩阵可获得每个样本点的逆近邻数,通过偏度可以衡量样本逆近邻数的非对称性,并以此分析数据集的hubness情况。

④Quick PCA降维

因为逆近邻的偏度与数据集的本征维数强烈正相关,因此将偏度作为主成分分析的降维指标,当偏度小于某一设定的阈值时便可认为数据集已损失了较多的本征维数,即剩下的维数为理想的k个主成分。为了加快此过程的搜寻速度,本章作了以下优化:首先将数据集的维数进行p等分并求出其对应的偏度,当该处偏度小于设定的阈值时停止运算;然后,针对此区间将这p等分的样本继续进行q等分,计算每一处的偏度直至该处偏度小于设定的阈值时停止运算,至此便可快速找到理想的k个主成分。

⑤聚类分析

由于hub可以代表局部中心性,{ 61% : 所以可以用hub聚类算法对降维后的数据进行聚类分析。 }

0. 算法思想

Quick PCA-Hub聚类算法首先对数据集进行预处理,将数据的每一维进行归一化;其次,构建KNN邻域矩阵,计算每个点的逆近邻数。然后,用PCA进行降维,在降维的过程中通过偏度的变化率来控制降维的程度,以防损失过多重要的有价值信息。最后,在获取降维数据后利用hub聚类算法进行聚类分析。下面是Q

uick PCA-Hub的聚类算法思想:

Algorithm. Qucik PCA-Hub.

```
float[][] knn=getKNN();
float[] Nk =getSkewness(knn);
float[][] eigenvectors=pca();
Dataset new=getKpca();
initializeClusterCenters();
Cluster[] clusters = formClusters();
float t = t0; initialize temperature
repeat
â€‹float = getProbFromSchedule(t);
â€‹for all Cluster c clusters do
â€‹if randomFloat(0,1) ^ then
â€‹DataPoint h = findClusterHub(c);
â€‹setClusterCenter(c, h);
â€‹else
â€‹for all DataPoint x c do
â€‹setChoosingProbability(x, );
â€‹end for
â€‹normalizeProbabilities();
```

â€¢DataPoint h = chooseHubProbabilistically(c);

â€¢setClusterCenter(c, h);

â€¢end if

â€¢end for

â€¢clusters = formClusters();

â€¢ = updateTemperature(t);

until noReassignments

return clusters

0. 实验结果及其分析

实验数据来源于加州大学尔湾分校(UCI)数据库,表4.1中第5列为真实数据集的偏度值,其中10代表近邻数k。虽然k值是固定的,但是使用其它的k值也可得到类似的结果。表4.1中第12列(迭代数)为搜索理想的k个主成分所需的次数,第13列(减少的维数)为数据集降维后所损失的维数。本章的Quick PCA-Hub算法分别与KMEANS、GHPKM、Ker-KM和Ker-GHPKM聚类算法进行比较,其中轮廓系数为聚类结果的评测指标,迭代数和减少的维数为搜索k个主成分的速度指标。实验结果如表4.1所示,下表中加粗的数据表示当前数据集的最优值。

表 4.1 UCI库中数据集的聚类质量(轮廓系数)

Table 4.1 Clustering quality expressed as silhouette index on data sets from the UCI repository.

数

据

集

样本数

维数

簇的个数

K-M

E

ANS

[9]

GH

PKM

[9]

Ker-KM [4]

Ker-GHP

KM [4]

PH-KM

Q PH-KM

迭代数

减少的维数

Ionosphere

351

34

2

1.72
0.28
0.28
0.28
0.25
0.41
0.40
8
11
mfeat-factors
2000
216
10
0.83
0.18
0.20
0.17
0.18
0.24
0.15
7
88
musk
2000
166
2
1.33
0.28
0.28
0.29
0.29
0.31
0.28
10
107
parkinsons
195
22
2
0.73

0.42
0.44
0.64
0.21
0.88
0.61
7
7
sonar
208
60
2
1.35
0.20
0.21
0.26
0.17
0.22
0.20
9
20
wpbc
198
33
2
0.86
0.16
0.16
0.32
0.22
0.31
0.51
8
11
AVG-UCI
0.25
0.26
0.33
0.22
0.39

0.36

8

41

下面主要从聚类结果的好坏和搜寻k个主成分的速度两个方面阐述Quick PCA-Hub聚类算法的聚类性能:

①聚类结果的好坏------轮廓系数

对于每一个数据集而言,取KMEANS、GHPKM、Ker-KM和Ker-GHPKM聚类算法中轮廓系数最大值作为经典聚类算法的最优值,然后同本章的Quick PCA-Hub聚类算法进行比较。{ 73% : 实验结果表明,相比之前的聚类算法, }本章提出的Quick PCA-Hub聚类算法在轮廓系数上提高了8%。{ 57% : 从表4.1和图4.2的实验结果可以看出, }经典的KMEANS聚类算法更适用于低维数据聚类;在数据集缺乏hubness特性的情况下,GHPKM、Ker-GHPKM等hub聚类算法表现不佳,其性能接近于KMEANS算法;然而当数据集呈现出较高的hubness特性时,GHPKM、Ker-GHPKM等hub聚类算法的表现要优于KMEANS算法。同时,本文提出的Quick PCA-Hub聚类算法无论数据集是否呈现出较高的hubness特性,均可以取得不错的聚类效果,相比之前的聚类算法适用范围更广,聚类性能更佳。从UCI平均数据集的轮廓系数可以观测到,本章的Quick PCA-Hub聚类算法要优于之前的聚类算法,但逊于第三章的PCA-Hub聚类算法。

图 4.2 PCA-Hub聚类算法对近邻数k的敏感程度

Figure 4.2 The sensitivity of PCA-Hub Clustering Algorithm to k nearest Neighbors.

②搜寻k个主成分的速度

图 4.3 PCA-Hub聚类算法对近邻数k的敏感程度

Figure 4.3 The sensitivity of PCA-Hub Clustering Algorithm to k nearest Neighbors.

从表4.1中第12列的迭代数和第13列的减少的维数可以看出,本章的Quick PCA-Hub在高维数据空间中搜索理想的k个主成分时表现出了巨大的优势,然而当数据集的维数不高时Quick PCA-Hub聚类算法的加速效果并不明显。同时可以看出,当数据集有较高的hubness特性时,Quick PCA-Hub聚类算法不仅加速搜寻k个主成分的速度,{ 60% : 而且可以获得更优的聚类结果; }然而当数据集有较低的hubness特性时,Quick PCA-Hub聚类算法的聚类优化效果则不明显。

快速搜索最近邻居

{ 75% : 最邻近搜索(Nearest Neighbor Search, NNS), }亦称为“最近点搜索”(Closest point search),是指在一个尺度空间里搜索最近点的最优化问题[49]。可以对问题进行如下描述:在尺度空间M中,存在一个点集S以及一个目标点,在点集S中找到离目标点q最近的点。最近邻搜索有多种解决方法,这些方法的性能取决于它们求解的时间复杂度以及搜索的空间复杂度。朴素最近邻搜索需要遍历整个点集,{ 59% : 计算目标点与其它点之间的距离, }并记录当前的最近点。朴素最近邻搜索较为初级,适用于较小规模的点集,但是对于较大尺度的点集和较高的空间维数并不适用。在最邻近搜索的几个变化中,最著名的是KNN(K-nearest neighbor algorithm)[50]和近似最邻近查找(-approximate nearest neighbor search)[51]。

Hub聚类算法需要计算KNN的完全图。由于未利用任何空间数据结构或近似计算的技术,朴素KNN图的计算复杂度在处理大型数据集时将会变得十分昂贵,但是可通过快速近似方法在合理的时间内构建一个十分精准的近似图。关于快速近似方法可以使Chen等人提出的通用方法[52]或者使用基于locality-sensitive hashing的特定度量近似方法[53]。不同的近似KNN搜索方法的性能取决于解决特定问题时数据集的数据特征和在特定环境下k近邻的本征难度[54]。近年来,人们开始关注在模糊和不确定数据集中随机逆k近邻查询的计算复杂度[55]。Tomaš̌ev等人针对一些hub聚类算法作了实验研究,实验结果表明即使在线性时间内构建KNN图也不会明显降低算法性能[2]。

本章小结

重庆大学硕士学位论文

重庆大学硕士学位论文

4 Quick PCA-Hub聚类算法

本章分别从快速搜索k个主成分和快速搜索最近邻居两方面增加PCA-Hub算法的聚类分析速度:首先,Quick PCA-Hub算法分别与经典聚类算法和PCA-Hub算法进行了对比分析。通过实验证明,Quick PCA-Hub算法相比经典聚类算法可以取得不错的聚类结果,而且当数据集的维数较高时,Quick PCA-Hub算法在搜索理想的k个主成分时表现出了巨大的优势。其次,从理论上探讨了快速搜索最近邻居的方法。虽然在大型高维数据数据空间中朴素KNN图的构建十分耗时,但是可以通过近似KNN搜索方法在合理的时间内构建一个十分精确的近

似图。

总结与展望

总结

{ 78% : 聚类分析被视为数据挖掘的一个十分重要的研究领域, }常常被用于分析现实生活的大量未知数据,从而发现其中重要的有价值信息和知识。{ 62% : 随着科技的发展,现实生活中未知数据的数量越来越多, }{ 58% : 聚类分析在实际应用中的地位也越来越重要。 }但同时数据集的尺度也越来越大,数据的维数也越来越高,这些问题将不断地向传统的聚类分析算法提出挑战。基于不同的理论以及聚类模型,{ 70% : 研究人员提出了多种聚类分析算法。 }针对高维数据空间中维数灾难这一问题,hub聚类算法利用高维数据空间的特征,可以解决传统聚类算法无法在高维数据空间中聚类分析的问题。

Hubness是最近几年才提出的一个较为新颖的概念,{ 65% : 常被应用于有监督的机器学习中, }例如分类和回归问题,而在无监督的机器学习中则研究不多。本文针对hubness这一概念,对将其应用到聚类分析中作了详尽的分析研究,{ 55% : 所得出的主要研究和结论如下: }

①详尽地概述了数据挖掘和聚类分析,分别从数据挖掘的定义、作用、发现过程以及应用等方面,{ 55% : 系统地对其进行了全面的介绍。 }接着阐述了聚类分析的定义及其主流的聚类模型,重点分析比较了常用的聚类分析算法,并归纳出了主流的聚类分析算法的适用范围及优缺点,同时介绍了聚类分析的评价标准和评估指标,并分析了聚类分析算法的现状和发展趋势。

②详尽地概述了维数灾难这一现象,并归纳出了在搜索高维空间数据时由于维数灾难可能导致的问题,本文深入地研究了其中的一个问题------hubness。针对hubness这一现象,首先对其进行了详细地描述并给出了形式化的定义,然后对其本质特征进行了仔细分析,包括表征的非对称性、hubs的在统计分布中的位置等等。根据这些特征进一步研究了hubs在聚类分析中的作用,并介绍了相关的hub聚类算法,同时归纳总结出了其各自的优缺点及适用范围。

③针对hub聚类算法未处理高维数据空间中的冗余和噪声特征,因此本文提出了PCA-Hub聚类算法。PCA-Hub聚类算法是以的偏度与本征维数强烈正相关为理论基础,通过构建数据集的KNN邻域矩阵,以偏度的变化率作为降维依据选出理想的k个主成分,之后再对降维后的数据集进行聚类分析。实验分别从聚类结果的好坏(轮廓系数)、对近邻数k的敏感程度和聚类结果的一致性三方面进行了分析,实验结果表明,无论数据集是否呈现出较高的hubness特性,PCA-Hub聚类算法均可以取得不错的聚类效果,相比之前的聚类算法,轮廓系数平均提高了15%;当数据集本身的维数较高时或者的偏度不低时,PCA-Hub聚类算法在使用不同近邻数k时表现出了相似的聚类性能,因此近邻数的选择对于PCA-Hub聚类算法的聚类结果影响并不强烈;{ 59% : 在实验环境和聚类算法参数一致的情况下, }PCA-Hub聚类算法的结果在很大程度上具有一致性。

④PCA-Hub聚类算法虽然可以很好地解决高维数据空间中的冗余和噪声数据,{ 68% : 然而随着数据集尺度和数据集维数的不断增加, }PCA-Hub聚类算法的耗时将会变得越来越严重甚至是不可接受。因此,本文分别从快速搜索k个主成分和快速搜索最近邻居两方面增加PCA-Hub算法的聚类分析速度。通过实验证明,Quick PCA-Hub算法相比经典聚类算法可以取得不错的聚类结果,而且当数据集的维数较高时,Quick PCA-Hub算法在搜索理想的k个主成分时表现出了巨大的优势。其次,从理论上探讨了快速搜索最近邻居的方法。虽然在大型高维数据数据空间中朴素KNN图的构建十分耗时,但是可以通过近似KNN搜索方法在合理的时间内构建一个十分精确的近似图,而且近似的KNN图并不会显著影响聚类结果。

展望

本人在高维数据空间中的聚类分析领域进行了一些研究并获得了一定的成果,但是由于本人的科学研究水平以及研究时间等因素的限制,论文中存在一些不足之处尚待改进以及尚未完成的研究,在未来的研究工作中还需要对以下的几个方面进行深入研究。

①PCA-Hub聚类算法在解决高维数据空间中的冗余和噪声数据时,{ 64% : 需要预先设定偏度下降的阈值。 }如果阈值设置的不合理,{ 61% : 那么聚类分析的结果可能就不理想。 }在今后的工作中,希望可以设置一个自适应的阈值来控制偏度下降的程度,从而降低PCA-Hub聚类算法对参数设置的敏感性。

②进一步探索不同的近似KNN搜索方法对PCA-Hub聚类算法的影响,以便可以找到一种合适的方法来在合理的时间内构建KNN图。

5 总结与展望

致谢

随着毕业时间的临近,{81% : 三年的研究生学习生涯也即将结束。 }回顾这几年的时光,有老师、同学以及朋友的亲切陪伴,一路走来,虽有辛劳,却也收获了成功的满足,值此,{ 62% : 向三年内关心帮助过我的老师、同学以及朋友们表示由衷的感谢。 }

{ 60% : 首先要向我的导师葛亮老师致以深深的谢意, }整个研究生阶段从入学进校到论文撰写准备毕业,{ 55% : 全程都贯穿着葛亮老师的热情关怀和悉心指导。 }三年来,葛老师多次加班加点帮我修改论文,多次为我的研究方案提出建议,无不使我获益匪浅,深受启发,并帮助我逐渐学会了如何去面对问题,思考问题最终圆满的解决问题。葛老师细致严谨、实事求是的作风态度都深深的影响了我,使我受益良多,{83% : 再次向葛老师致以我最诚恳的谢意。 }

三年研究生阶段中,课题组朱庆生老师、舒立春老师、张志劲老师、胡琴老师在学习生活中给予了我极大的关心和鼓励,此外实验室的廖瑞金老师,王有元老师,李剑老师,杨丽君老师,杨庆老师,袁涛老师等也在各方面给予了我诸多帮助,在此向他们致以由衷的谢意!

然后还要真诚感谢已毕业的吴尧师兄,从进校的时候手把手指导我完成试验,其后耐心指导我进行科技论文写作,以及毕业后仍不忘跟我交流学习、科研以及工作心得,吴尧师兄给予我的帮助实在难以用言语表达;此外孙晓峰,刘健两位师兄也对我科研、生活给予了无私帮助,在此一并感谢。

另外我取得的科研成绩也离不开同窗许可的大力协助以及杨洪椿,石璧,钟睿,李洋洋等师弟们的配合。而在科研期间,与赵洪彬,陈勇,白洋,王慕宾,范才进,罗异等各位同学的交流也给予我诸多启发。此外实验室的各位同学也给予我诸多指导和帮助,{ 75% : 在此也向他们表达我最真诚的谢意。 }另外好友余瑜,彭珊,陈祥等也在我低谷时给予我诸多鼓励和帮助,在此一并表示感谢。

{ 61% : 谨以此文献给辛勤养育我的父母, }{ 55% : 感谢他们二十多年来的无私付出以及对我求学的支持,并为我创造了良好的学习,生活条件, }{ 58% : 在此,向他们致以我最崇高的敬意! }

最后,{ 78% : 衷心感谢在百忙之中抽出时间来评阅我论文和答辩的各位专家、教授! }

郎江涛

二零一七年四月于重庆

致 谢

参考文献

[1] John,N.,Megatrends:Ten New Directions Transforming Our Lives. NY:Futura,1984:p. 28.

[2] Milosˇ Radovanovic Ì•,Alexandros Nanopoulos,MirjanaIvanovic Ì•. Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data[J],Journal of Machine Learning Research 11 (2010) 2487-2531. 2010

[3] John,N.,Megatrends:Ten New Directions Transforming Our Lives. NY:Futura,1984:p. 28

[4] Data Mining Curriculum. ACM SIGKDD. 2006-04-30.

[5] 潘有能,XML 挖掘:聚类、分类与信息提取. 杭州:浙江大学出版社,2012

[6] Lloyd, S. (1982). "Least squares quantization in PCM". IEEE Transactions on Information Theory. 28 (2): 129–137

[7] Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise". In Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231.

[8] Ankerst, Mihael; Breunig, Markus M.; Kriegel, Hans-Peter; Sander, Jörg (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". ACM SIGMOD international conference on Management of data. ACM Press. pp. 49–60.

[9] Roy, S.; Bhattacharyya, D. K. (2005). "An Approach to find Embedded Clusters Using Density Based Techniques". LNCS Vol.3816.

[10] Cheng, Yizong (August 1995). "Mean Shift, Mode Seeking, and Clustering". IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE. 17 (8): 790–799

[11] Xu, X.; Yan, Z.; Xu, S. (2015). "Estimating wind speed probability distribution by diffusion-based kernel density method". Electric Power Systems Research. 121: 28–37.

[12] Sculley, D. (2010). Web-scale k-means clustering. Proc. 19th WWW.

[13] Huang, Z. (1998). "Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery. 2: 283–304.

[14] R. Ng and J. Han. "Efficient and effective clustering method for spatial data mining". In: Proceedings of the 20th VLDB Conference, pages 144-155, Santiago, Chile, 1994.

[15] Tian Zhang, Raghu Ramakrishnan, Miron Livny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD, pp. 103–114.

[16] McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178

[17] Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases". ACM Transactions on Database Systems. 15 (4): 483–517.

[18] Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to Information Retrieval. Cambridge University Press. ISBN 978-0-521-86571-5.

[19] Estivill-Castro, Vladimir (20 June 2002). "Why so many clustering algorithms — A Position Paper". ACM SIGKDD Explorations Newsletter. 4 (1): 65–75.

[20] Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to Information Retrieval. Cambridge University Press. ISBN 978-0-521-86571-5

[21] Färber, Ines; Günnemann, Stephan; Kriegel, Hans-Peter; Kröger, Peer; Müller, Emmanuel; Schubert, Erich; Seidl, Thomas; Zimek, Arthur (2010). "On Using Class-Labels in Evaluation of Clusterings" (PDF). In Fern, Xiaoli Z.; Davidson, Ian; Dy, Jennifer. MultiClust: Discovering, Summarizing, and Using Multiple Clusterings. ACM SIGKDD

[22] Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical Association. American Statistical Association. 66 (336): 846–850

[23] Powers, David M W (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). Journal of Machine Learning Technologies. 2 (1): 37–63.

[24] Oommen, T. .; Misra, D. .; Twarakavi, N. K. C.; Prakash, A. .; Sahoo, B. .; Bandopadhyay, S. . An Objective Analysis of Support Vector Machine Based Classification for Remote Sensing. Mathematical Geosciences. 2008, 40 (4): 409. doi:10.1007/s11004-008-9156-6.

[25] Hughes, G.F., 1968. "On the mean accuracy of statistical pattern recognizers", IEEE Transactions on Information Theory, IT-14:55-63.

[26] Not to be confused with the unrelated, but similarly named, Hughes effect in electromagnetism (named after Declan C. Hughes) which refers to an asymmetry in the hysteresis curves of laminated cores made of certain magnetic materials, such as permalloy or mu-metal, in alternating magnetic fields.

[27] Groeneveld, RA; Meeden, G. Measuring Skewness and Kurtosis. The Statistician. 1984, 33 (4): 391–399

[28] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. On the surprising behavior of distance metrics in high dimensional spaces. In Proceedings of the 8th International Conference on Database Theory (ICDT), volume 1973 of Lecture Notes in Computer Science, pages 420– 434. Springer, 2001

[29] Maritz. J.S. (1981) Distribution-Free Statistical Methods, Chapman & Hall. ISBN 0-412-15940-6. (page 217)

[30] Myers, Jerome L.; Well, Arnold D., Research Design and Statistical Analysis 2nd, Lawrence Erlbaum: 508, 2003, ISBN 0-8058-4037-0

[31] Kriegel HP, Kr¨oger P, Zimek A (2009) Clustering high-dimensional data: A survey on subspaceclustering, pattern-based clustering, and correlation clustering. ACM Transactions on KnowledgeDiscovery from Data 3(1):1:1–1:58

[32] Jing L, Ng M, Xu J, Huang J (2005) Subspace clustering of text documents with feature

weightingk-means algorithm. In: Ho T, Cheung D, Liu H (eds) Advances in Knowledge Discovery andData Mining, Lecture Notes in Computer Science, vol 3518, Springer Berlin Heidelberg, pp802–812

[33] Li T, Ma S, Ogihara M (2004) Document clustering via adaptive subspace iteration. In: Proceedingsof the 27th Annual International ACM SIGIR Conference on Research and Developmentin Information Retrieval, ACM, New York, NY, USA, SIGIR ' 04, pp 218–225, DOI

[34] Jing L, Ng M, Huang J (2007) An entropy weighting k-means algorithm for subspace clusteringof high-dimensional sparse data. Knowledge and Data Engineering, IEEE Transactions on19(8):1026–1041

[35] Nenad Toma sev,Milo s Radovanovi c,DunjaMladeni c,andMirjana Ivanovi c. The Role of Hubness in Clustering High-Dimensional Data[J],IEEETRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 3,2014

[36] Amina M,Syed Farook K. A Novel Approach forClustering High-Dimensional Data using Kernel Hubness[J]. InternationalConfenrence on Advances in Computing and Communication. 2015.

[37] D. Corne, M. Dorigo, and F. Glover, New Ideas in Optimization. McGraw-Hill, 1999.

[38] Grigorios F. Tzortzis and Aristidis C. Likas, The Global Kernel-Means Algorithm forClustering in Feature Space IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 20, NO. 7, JULY 2009

[39] D. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data" Proc Natl Acad Sci U S A. 2003 May 13; 100(10): 5591–5596

[40] Rosman G., Bronstein M. M., Bronstein A. M. and Kimmel R., Nonlinear Dimensionality Reduction by Topologically Constrained Isometric Embedding, International Journal of Computer Vision, Volume 89, Number 1, 56–68, 2010

[41] Bennett, R. (June 1965). "Representation and analysis of signals—Part XXI: The intrinsic dimensionality of signal collections". Rep. 163 (PDF). Baltimore, MD: The Johns Hopkins University.

[42] Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002, XXIX, 487 p. 28 illus. ISBN 978-0-387-95442-4

[43] Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space (PDF). Philosophical Magazine. 1901, 2 (6): 559–572

[44] Rolewicz, Stefan (1987), Functional analysis and control theory: Linear systems, Mathematics and its Applications (East European Series), 29 (Translated from the Polish by Ewa Bednarczuk ed.), Dordrecht; Warsaw: D. Reidel Publishing Co.; PWN—Polish Scientific Publishers, pp. xvi+524, ISBN 90-277-2186-6, MR 920371, OCLC 13064804

[45] Deza, Elena; Deza, Michel Marie (2009). Encyclopedia of Distances. Springer. p. 94.

[46] For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution; See Donoho, David L (2006). "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution". Communications on pure and applied mathematics. 59: 797–829. doi:10.1002/cpa.20132

[47] Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013

[48] Peter J. Rousseeuw. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis[J]. Computational and Applied Mathematics. 20: 53–65. 1987.

[49] Roussopoulos, N.; Kelley, S.; Vincent, F. D. R. (1995). "Nearest neighbor queries". Proceedings of the 1995 ACM SIGMOD international conference on Management of data – SIGMOD '95. p. 71. doi:10.1145/223784.223794. ISBN 0897917316.

[50] Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185. doi:10.1080/00031305.1992.10475879.

[51] Ma, Zongmin. Artificial Intelligence for Maximizing Content Based Image Retrieval. IGI Global. p. 135. ISBN 9781605661759.

[52] Chen J, ren Fang H, Saad Y (2009) Fast approximate kNN graph construction for high dimensional data via recursive Lanczos bisection. Journal of Machine Learning Research 10:1989–2012

[53] Satuluri V, Parthasarathy S (2012) Bayesian locality sensitive hashing for fast similarity search.Proc VLDB Endow 5(5):430–441

[54] He J, Kumar S, Chang SF (2012) On the difficulty of nearest neighbor search. In: International Conference on Machine Learning (ICML), icml.cc / Omnipress

[55] Zhang P, Cheng R, Mamoulis N, Renz M, Zufle A, Tang Y, Emrich T (2013) Voronoi-based nearest neighbor search for multi-dimensional uncertain databases. In: Data Engineering (ICDE), 2013 IEEE 29th International Conference on, pp 158–169, DOI 10.1109/ICDE.2013.6544822

[56] Pyle, D., 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California.

重庆大学硕士学位论文

参考文献

附 录

A.作者在攻读硕士学位期间撰写的论文目录

[1] 胡建林,蓝彬桓,许 可,蒋兴良,石 璧,杨洪椿,风机叶片运用疏水性涂层防覆冰的风洞试验研究,高电压技术 (EI期刊,已录用待刊出,稿件编号:20160025)

[2]

B.作者在攻读硕士学位期间参与的科研项目

[1]