学校编码：10384　　分类号_____密级_____
学号：24320111152272　　UDC_____

厦门大学

工　学　硕　士　学　位　论　文

基于聚类分析的可视化技术及其应用研究

Research on Visualization Technology and Its Application Based on Clustering Analysis

蔡朱华

指 导 教 师：董　槐　林　教　授

专 业 名 称：计算机软件与理论
论文提交日期：２０１４　年　５　月
论文答辩日期：２０１４　年　５　月
学位授予日期：２０１４　年　６　月

指　导　教　师：_____

答辩委员会主席：_____

2014 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果,均在文中以适当方式明确标明,并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名）:

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　√　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘　要

随着科技的进步特别是整个信息产业的快速发展，我们的社会进入了一个崭新的信息时代。不仅数据采集能力和手段越来越多样化，存储设备技术也迅猛发展，数据采集与存储设备的不断发展带来了大数据的时代。面对大量且繁杂的数据信息，如何从中提取出有价值且便于用户观察的信息是最迫切而重要的问题。显然要解决上述的问题，仅仅采用数据挖掘容易造成得到的信息不易被理解或不一定正确的问题，因而本文研究可视化数据挖掘技术，将数据挖掘和数据可视化技术结合在一起，相辅相成。

目前可视化技术与数据挖掘算法的联系是较松散，针对这一现象，本文主要研究内容是如何将数据挖掘算法与可视化技术更好地更高效地融合，并以聚类分析算法为切入点将挖掘过程中的数据可视化、过程可视化及结果可视化进行相应的研究，同时也提供在社交网络、科学研究领域等方面的应用示例。主要研究内容如下：

（1）提出了一种基于 MASI 距离的层次聚类方法，并融合了随机抽样的方法，对层次聚类算法进行了改进并将算法应用到职业网络数据集中实现了结果可视化。采用了随机抽样之后的层次聚类算法的时间复杂度有效地降低了，并且在聚类结果可视化以不同的树形图进行呈现，一目了然。

（2）提出基于 SOM 聚类的可视化模型，并将模型应用到大气温度数据集实现聚类过程及聚类结果的可视化，并提出了基于聚类的兴趣度量和基于近邻的兴趣度量来对属性进行排名，优化了数据挖掘结果的可视化。值得一提的是该应用中的交互可视化设计，它结合了颜色映射、缩放等交互技术让用户更方便地进行观察和分析数据。

（3）将平行坐标可视化技术与 K-Means 算法相结合，在实验过程中通过可视化技术同时对数据和挖掘结果进行可视化，从而提高了算法的效率和准确度。以 Iris 数据集为测试数据对 K-Means 算法可视化的有效性进行验证，实验表明相较于传统的 K-Means 算法，其效率和正确率都有较大的提高。

关键词：可视化技术；可视化数据挖掘；聚类分析

Abstract

With the rapid development of science and technology especially information industry, our society has entered a new era of information. Not only the data collection capacity and means become more and more diversified but also the storage device technology is getting better and better. The continuous development of data acquisition and storage devices has brought the era of big data. Facing a great deal of data and complex information, how to extract valuable information and making it easily understood for users is the most urgent and important issue. Using the data mining is likely to cause the questions that the information is not easy to be understood or not right. So to solve the above problem just using data mining is not enough. Visual data mining is proposed in this thesis. We study visual data mining technology that combines data mining and data visualization technology together.

The combination of visualization technology with data mining algorithm is relatively loose. Aiming at this phenomenon, main research content of this thesis is how to integrate data mining algorithm and the visualization technology better and more efficiently. Clustering analysis algorithm is chose as the breakthrough point of the research of the data visualization, the visualization of process and the visualization of result. Applications of social network and scientific research are provided. The main research contents are as follows:

(1) A hierarchical clustering method based on MASI distance which integrates random sampling method is proposed in this thesis. The hierarchical clustering algorithm was improved. The algorithm is applied to the professional network data set and the results are visualized. After adopted random sampling in the hierarchical clustering algorithm, the time complexity of the algorithm effectively reduces. The clustering results are visualized in different tree diagram, be clear at a glance.

(2) The visual model based on SOM clustering is put forward. The model is applied to the atmospheric temperature data set to realize the clustering process visualization. The interest measurement based on clustering and interest based on

neighbor metrics are proposed to rank attributes, optimizing the visualization of the data mining results. The interactive visual design of this application is worth to be mentioned. The design uses the technologies such as color mapping and scaling to allow users more easily to observe and analyze the data.

(3) The parallel coordinate visualization technology is combined with K-Means algorithm. Efficiency of the algorithm is improved by visualization technology. In the experiment both data and results is visualized. The improved K-Means algorithm is tested by Iris data set. The experiment proves that compared with the traditional K-Means algorithm, the efficiency and accuracy of the improved K-Means are better.

Key Words: Visualization; Visual Data Mining; Clustering Analysis

目 录

Contents

第一章 绪论

1.1 研究背景及意义

随着科技的进步特别是整个信息产业的快速发展，我们的社会进入了一个崭新的信息时代。计算机的不断普及以及数据库技术的不断成熟带来了数据库管理系统的广泛应用。

伴随着数据形式的层出不穷和多样化，数据采集能力和手段也越来越多样化。一方面随着条形码和信用卡在商业、金融、保险等领域广泛使用，基于射频识别技术的数据采集方式应运而生，美国零售商 WalMart（沃尔玛）每天通过条形码读入器获取 3 亿条左右的交易数据。另一方面，物联网技术的应用和普及也带来了数据采集能力的不断增强，物联网是物物相连的互联网，它通过数据采集设备把物品与互联网进行连接从而实现对物品的智能化识别、定位、监控以及管理，这些数据采集设备包括红外感应器、全球定位系统、激光扫描器等。各行各业的信息化进程都加速了数据采集的发展。

另一方面随着存储设备与技术的迅猛发展，大容量、高速度、低成本的存储设备和系统已经相继问世。数据采集与存储设备的不断发展带来了大数据的时代。但大数据在给人们带来方便的同时也带来了一系列的问题比如信息量过大以至于超过了人们掌握消化信息的能力；一些信息因为难辨真伪而给信息的正确性判断带来了困难；网络信息难以保障安全的问题；数据形式的多样性带来了信息进行统一处理的困难。同时人们意识到隐藏在大数据后更深层次和更重要的能够描述数据整体特征的信息是可以预测发展趋势，这在进行专家决策时提供了重要的参考价值。面对大量且繁杂的数据信息，如何从中提取出有价值且便于用户观察的信息是最迫切而重要的问题。

显然要解决上述的问题，如果仅仅采用数据挖掘进行解决，则容易得到复杂模糊的信息甚至错误的信息，造成了信息的不易理解和结果正确率低的问题。因此本文提出采用数据可视化技术与数据挖掘相结合的方式即所谓的可视化数据挖掘技术。

传统的数据挖掘过程是以算法为中心的，在挖掘的过程中通常是没有用户的参与，数据挖掘过程的不智能问题、挖掘结果复杂难懂问题接踵而至。在这一点上结合了可视化技术的数据挖掘是以人的交互为中心的。在过程中用户可以有效地参与进来——用户可通过交互手段来调用挖掘算法、在过程中用户可实时观察挖掘出的信息并做出相应的判断、纠正等交互，最后的结果以可视化形式展示，从而提高了整个数据挖掘过程的灵活性、交互性以及有效性。

可视化数据挖掘将数据挖掘与可视化紧密地联系在了一起，对可视化数据挖掘技术的研究具有重要的意义：

（1） 扩展了数据的表达力和理解力，利用人类对模式的识别能力剖析数据存在的规律，从而揭示数据中的内在本质和联系。由于人类对复杂数据的认知度较低，对发现数据内部的规律有一定的局限。而可视化数据挖掘可通过图形和可视化工具在数据准备阶段对数据进行呈现，如此一来，用户可在图形模式下更容易找到数据中可能存在的模式、关系甚至异常等。

（2） 可视化技术是用户与数据挖掘系统的交互桥梁，用户的交互参与使得专家领域的知识能更好的规整和约束挖掘过程，改善挖掘结果。目前，大部分的数据挖掘都有结合一些可视化方法，但更多的仅是对数据而不是过程进行可视化，这样的结合是松散且挖掘效果不佳。然而可视化数据挖掘是对数据挖掘的过程和可视化技术进行紧密结合，从而使得用户可适时地根据可视化交互引导挖掘算法的选择、监控和引导挖掘过程和中间结果。这种融合了用户决策的挖掘机制不仅使得挖掘过程更为灵活也完善合理化了挖掘的结果。

（3） 通过对挖掘结果的直观的可视化使用户更易于发现和分析数据，用户可从直观的可视化图形中发现潜在的有用信息。

随着数据挖掘研究的发展，迫切需要将数据挖掘与可视化相结合。在数据挖掘过程中，从数据预处理、挖掘过程和挖掘结果的可视化对数据挖掘的交互性、友好性都有很重要的意义。

1.2 国内外研究现状

1.2.1 数据挖掘及其现状

数据挖掘(Data Mining)起源于 KDD（Knowledge Discovery in Database，数据库中的知识发现），可追溯到 20 世纪 80 年代末。1995 年第一届知识发现和数据挖掘国际学术会议在加拿大召开，在会议上数据被形象地比喻为矿床，对“数据库中的知识发现”的形象描述——“数据挖掘”就迅速地传播开来[1]。

从严格意义上说，数据挖掘虽然只是数据库中的知识发现中的一个步骤，但它却是最为举足轻重的一个步骤。数据挖掘的定义从最初简单的描述扩展到现在较为复杂多样的定义，一种被广泛接受的定义是数据挖掘是一个从不完整的、不明确的、大量的并且包含噪声，具有很大随机性的实际应用数据中，提取出隐含其中、事先未被人们获知、却潜在有用的知识或模式的过程[2]。

数据挖掘是一个融合了多学科如人工智能、数据库技术、机器学习、统计学以及信息检索等的研究领域，经过这十几年来的研究发展在该研究领域产生了很多新概念和方法，但也因此一些基本的概念和方法也慢慢成熟和清晰。数据挖掘研究正朝着更深入的多学科交叉方向发展[3]。

在大数据时代的背景下，由于传统的数据挖掘算法在运行时间和空间的限制已经不能很好地适应数据量的剧增和数据形式的变化。因此，高性能的大规模数据集的挖掘算法的研究已经成为了重要的研究领域。目前，国内外研究员不断地对传统数据挖掘算法进行改进和优化并取得了较为突出的成绩，尤其是在数据挖掘聚类这一重要领域。

在大规模数据下数据挖掘聚类这一研究领域有国外很多研究员做了大量的改进及优化工作。针对大规模稀疏数据集 R.Ng 等提出了 CLARA 算法[4]以及 M.Ester 等提出了 DBSCAN[5]算法都是有效进行聚类挖掘的算法。微软研究院在 K-means 算法基础上根据大数据聚类架构提出了 Scalable-kmeans[6]算法，该方法由于高效存储性能被集成到 SQL SERVER 中。研究人员同样在分布式环境下也提出了很多的并行聚类算法比如 Sanpawat Kantabutra 等改进的并行 k-means 算法[7]。E.Januzaj 等提出了基于密度的分布式聚类算法 DBDC[8]、适用于空间数据的

DBDC 的改进算法 SDBDC[9]。

在认识到大数据集的挖掘聚类算法的重要性，国内很多研究所、高效及科研机构也进行了大量的研究工作，取得了一系列的进展。复旦大学的周水庚等人对 DBSCAN 进行改进提出了一种快速高效的聚类的算法 FDBSCAN[10]；吉首大学的段明秀将 CLARA 算法与自组织特征映射算法（SOFM）相结合提出一种新型的聚类算法[11]。对于分布式环境下的聚类，国内科研人员也有所成就如利用 Google 的 Hadoop 实现了基于云计算的聚类算法[12,13]。可以看出虽然国内在大规模数据挖掘聚类分析领域也做了大量的研究，但相对国外研究情况来说国内数据挖掘起步较晚，尚需要更为广泛和更深入的研究。

数据挖掘同时也是一门面向应用的学科，研究人员将工作重心放在挖掘算法的研究改进并将其运用到实际应用中去。随着数据挖掘的发展，相应的数据挖掘应用也在不断发展成长着。人们针对特定领域的应用开发了许多专用的数据挖掘工具包括客户关系管理、气象、医学、生物医学、物理研究、金融、零售业和电信业的数据挖掘工具等。以上这些将数据挖掘与领域知识相结合的应用提供了满足特定需求的数据挖掘解决方案。

国内外很多企业研究所都着力于研究数据挖掘系统。由加州理工学院喷气推进实验室研制出 SKICAT（Sky Image Cataloging and Analysis Tool）工具可以辅助天文学家去发现新的类星体的工具；IBM 公司为通信行业研发了一整套的商业智能解决方案 Intelligent Miner，为用户提供了从市场分析、客户分析、客户关系管理到综合决策分析的全方面的技术支持；而 Marksman 是一款美国 Firstar 银行使用的数据挖掘工具，它可以根据消费者的理财方式进行分类并进一步预测何时该向相应的客户提供什么样的理财产品。数据挖掘的研究正朝着更实用的技术应用方向发展。

数据挖掘的研究现状虽然较为积极乐观，但仍存在着一些问题和挑战，仍需要进行更深入更广泛的研究。主要表现在几个方面：

（1）数据挖掘技术跟不上数据存储类型的变化，目前为止数据挖掘技术较为单一化，而数据存储类型却越来越多样化。单一的数据挖掘技术是不可能适合多样的数据存储类型的，如何根据特定的数据类型制定与之相应的数据挖掘技术