国家自然科学基金青年项目（61402482）
中国博士后基金特别资助项目（2015T80555）
项目（1501012A）
BY2015023-05）

复杂多源数据的知识获取与知识发现

周　勇　著

中国矿业大学出版社
China University of Mining and Technology Press

青年项目(61402482)
中国博士后基金特别资助项目(2015T80555)
江苏省博士后科研资助计划项目(1501012A)
江苏省产学研前瞻性项目(BY2015023－05)

复杂多源数据的知识获取与知识发现

周　勇　著

中国矿业大学出版社

内容简介

本书针对复杂多源数据的知识获取与知识发现问题，分别从模式识别中聚类和分类的角度进行研究，并以复杂网络的知识发现为例，研究了三种改进的复杂网络社区发现方法。本书是关于知识获取和知识发现方法的专著，内容新颖、图文并茂、立足应用，可供机器学习、模式识别、计算机应用技术等领域的技术人员阅读，也可作为相关学科的培训教材。

图书在版编目(CIP)数据

复杂多源数据的知识获取与知识发现/周勇著.
徐州：中国矿业大学出版社，2015.10
ISBN 978-7-5646-2801-7

Ⅰ.①复… Ⅱ.①周… Ⅲ.①知识获取－研究 Ⅳ.①TP18

中国版本图书馆 CIP 数据核字(2015)第 200433 号

书　　名　复杂多源数据的知识获取与知识发现
著　　者　周　勇
责任编辑　仓小金
出版发行　中国矿业大学出版社有限责任公司
（江苏省徐州市解放南路　邮编 221008）
营销热线　(0516)83885307　83884995
出版服务　(0516)83885767　83884920
网　　址　http://www.cumtp.com　**E-mail**：cumtpvip@cumtp.com
印　　刷　江苏徐州新华印刷厂
开　　本　787×960　1/16　**印张** 7.75　**字数** 147 千字
版次印次　2015 年 10 月第 1 版　2015 年 10 月第 1 次印刷
定　　价　26.00 元

前　言

近年来，随着信息技术和数据库技术的迅猛发展，人们可以非常方便地获取和存储大量的数据，面对大规模的海量数据，人们急需高效的知识获取与知识发现方法。本书针对复杂多源数据的知识获取与知识发现问题，分别从模式识别中聚类和分类的角度进行研究，并以复杂网络的知识发现为例，研究了三种改进的复杂网络社区发现方法。全书共分为9章，主要内容包括：

(1) 基于序号编码的改进遗传模糊聚类算法

针对已有的基于遗传算法的FCM聚类算法计算复杂度高的缺点，在已有遗传算法与FCM算法相结合的算法的基础上对编码方式进行了改进，提出了基于序号编码的改进GFCA算法。用两组标注数据集对提出的算法进行了仿真实验，实验结果证明了该算法能在不损失聚类准确率的前提下减少计算复杂度。

(2) 基于弧度点对称距离的自适应动态聚类算法

该算法结合遗传算法和聚类算法各自的优点，采用动态变长编码方式，实现自动演化聚类数目。在本算法中，利用弧度点对称距离取代欧几里得距离，将点分配到不同的聚类中，既能够发现凸面的聚类又能够发现非凸面的聚类，而且无论聚类的大小和形状如何，只要聚类拥有对称特性，该算法都能适用。本算法充分发挥了遗传算法的全局寻优能力和一般聚类算法的局部搜索能力，可以更好地提高聚类质量。

(3) 基于自适应遗传算法的有趣分类规则知识发现

在数据挖掘中数据分类是一个很重要的研究方面，然而现有的

成熟的分类算法一般只是从数据库中发现准确度很高的规则，而对发现有趣分类规则却论述不多，针对这一问题，提出了一种基于遗传算法挖掘有趣分类规则的方法。首先，通过属性信息增益和设置属性信息增益权值以及规则的准确度来构造有趣分类规则的适应度函数，实现了对有趣规则的客观评价和主观评价的统一；其次，为了防止进化过程过早收敛或降低收敛速度，使用了自适应的遗传算法。最后使用 JBuilder 2006 平台给出了实验结果，从数据库中发现了有趣的分类规则，验证了算法的有效性。

(4) 基于无参核学习的 Laplacian 正则化最小二乘分类

针对 Laplacian 正则化最小二乘分类算法中人为指定核函数类型导致算法性能难以达到最优的问题，提出一种基于无参核学习(Non-Parametric Kernel Learning，NPKL)的 NPKL-LapRLSC 算法。NPKL 算法直接从数据中学习能更好地描绘数据相似性特征的半正定核矩阵，有效避免了针对特定问题选择和构造合适的核函数、根据实际样本数据确定核函数的参数等难题，提高了 LapRLSC 算法在实际复杂应用中的适应能力。实验结果表明，选取合适的核函数对分类效果有很大的影响，而本方法具有较高的识别率，且更为高效，更具有扩展性。

(5) 复杂网络的社区发现方法

社区发现是近年来社会网络中的研究热点。针对近邻传播算法的偏向参数影响聚类结果的类簇，提出了一种基于自适应近邻传播的社区发现算法，该算法将模块度引入到算法迭代过程中，作为评价指标引导算法向最优的聚类结果运行，该算法同 AP 算法一样不需要用户设定输入参数，不需要预先指定社区数量。通过引入标签传播算法的思想，提出了一种基于改进标签传播的社区发现算法，引入局部度中心的概念，首先找出网络中所有局部度中心节点，给它们以及其各自的邻居节点分配一个唯一的标签，然后进行异步标签更新。为了提高复杂网络中社团检测的性能，提出了一种基于结构相似度仿射传播的社团检测算法。该算法首先选取结构相似度作为节点之

间的相似性度量，并采用了一种优化的方法来计算相似度。其次将计算得到的相似度矩阵作为输入，采用快速仿射传播算法进行聚类，得到最终的社团结构。实验结果表明，改进的算法在绝大多数情况下都得到了优于原算法的结果，与其他经典社区发现算法相比，改进的社区发现算法也是有效的。

本书是作者近年来在数据挖掘和机器学习研究成果的基础上总结而成的。值此著作完成之际，感谢课题组夏士雄教授、孟凡荣教授的关心和支持，感谢王志晓副教授在本书出版过程中给予的热情帮助。本书内容的完成，得到了朱牧、邢艳、刘蓓佐、李佑文、周然然、石梦雨、孙贵宾等研究生的大力支持和帮助，在此一并表示感谢。

由于作者水平所限，书中不足之处在所难免，恳请读者批评指正。

著　者

2015 年 5 月

目　录

1 绪 论

1.1 研究背景与意义

随着信息技术和数据库技术的迅猛发展，人们可以非常方便地获取和存储大量数据。面对海量的数据，传统的数据分析工具（如管理信息系统）只能进行一些表层的处理（如查询、统计等），而不能获得数据之间的内在关系和隐含的信息。为摆脱“数据丰富，知识贫乏”的困境，人们迫切需要一种能够智能自动地将数据转换为有用信息和知识的技术和工具，这种对强有力数据分析工具的迫切需求使得数据挖掘技术应运而生[1-3]。

数据挖掘（Data Mining），又称为数据库中的知识发现（KDD），是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含的、未知的、有潜在应用价值的信息或模式的过程。它是一门新兴的交叉学科，汇集了来自数据库技术、统计学、机器学习、高性能计算、模式识别、神经网络、数据可视化、信息检索、图像与信号处理和空间数据分析等各领域的研究成果[4,5]。

聚类是一种重要的数据分析技术，它搜索并识别一个有限的种类集合或簇集合，从而描述数据。聚类分析作为统计学的一个分支，已被广泛研究了许多年。聚类分析也已广泛地应用到许多领域中，包括模式识别、数据分析、图像处理以及市场研究。通过聚类，人们能够识别密集的和稀疏的区域，从而发现全局的分布模式以及数据属性之间的相互关系。在商务上，聚类能帮助市场分析人员从客户

基本信息库中发现不同的客户群，并且用购买模式来刻画不同客户群的特征。在生物学上，聚类能用于推导植物和动物的分类，对基因进行分类，获得对种群中固有结构的认识。聚类在地球观测数据库中相似地区的确定，汽车保险单持有者的分组，根据房屋的类型、价值和地理位置对一个城市中房屋的分组上也可以发挥作用[6]。

聚类分析是研究数据挖掘技术的有效手段，是一种无监督的分类方法。聚类的目标是把一个无类别标记的数据集按某种准则划分为不同的簇，使相同簇中数据相似性尽可能大，而不同簇间数据相似性尽可能小。聚类可分为基于层次的聚类、基于划分的聚类、基于密度的聚类、基于模型的聚类、基于网格的聚类等[7]。

作为一个数据挖掘的重要功能，聚类分析能作为一个独立的工具来获得数据的分布情况，观察每个类的特点，集中对特定的某些类进行进一步的分析，如对 Web 上的文档进行分类以发现信息。此外，聚类分析也可以作为其他算法（如关联分析和分类）的预处理步骤，这些算法再在生成的类上进行处理，可以大大提高这些算法的执行效率。因此聚类分析已经成为数据挖掘中一个非常重要的研究领域，具有重要的理论意义和实际应用价值。

1.2 国内外研究现状

将物理或抽象对象的集合分成由类似对象组成的多个类的过程被称为聚类（Clustering）。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个类簇中的对象彼此相似，与其他类簇中的对象相异[6]。

聚类分析起源于分类学，但聚类不等于分类。聚类与分类的不同在于聚类所要求划分的类是未知的，进行聚类前并不知道将要划分成几个组和什么样的组，也不知道根据哪些空间区分规则来定义组。它的目的是使得属于同一个簇的样本之间应该彼此相似，而不同簇的样本应该足够不相似。

聚类分析又称群分析，它是研究(样品或指标)分类问题的一种统计分析方法。聚类分析也称无监督学习或无指导学习，聚类的样本没有标记，需要由聚类学习算法来自动确定；在分类中，对于目标数据库中存在哪些类是知道的，要做的就是将每一条记录分别属于哪一类标记出来[6-9]。

目前，文献中存在很多种聚类分析的算法，而算法的选择取决于数据的类型、聚类的目的和应用。如果聚类分析被用作描述或探测的工具，可以对同样的数据采用多种算法以发现数据揭示的结果。现在主要的聚类算法大致可分为以下几种：基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法以及基于模型的方法[8-11]。

(1) 基于划分的方法

基于划分的方法是指给定一个有 N 个元组或者记录的数据集，基于划分的方法将构造 K 个分组，每一个分组就代表一个聚类，$K<N$。而且这 K 个分组满足下列条件：① 每一个分组至少包含一个数据纪录。② 每一个数据记录属于且仅属于一个分组(这个要求在某些模糊聚类算法中可以放宽)；对于给定的 K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都较前一次好，而所谓好的标准就是：同一分组中的记录越近越好，而不同分组中的纪录越远越好。使用这个基本思想的算法有：K-MEANS 算法、K-MEDOIDS 算法、CLARANS 算法[12-20]。

(2) 基于层次的方法

基于层次的方法对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。例如在“自底向上”方案中，初始时每一个数据纪录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。代表算法有：BIRCH 算法、CURE 算法、CHAMELEON 算法等[21-28]。

(3) 基于密度的方法

基于密度的方法(Density-based Methods)与其他方法的根本区别是:它不是基于各种距离的,而是基于密度的。这样就能克服基于距离的算法只能发现"类圆形"的聚类的缺点。这个方法的指导思想是只要一个区域中的点的密度大过某个阈值,就把它加到与之相近的聚类中去。代表算法有:DBSCAN 算法、OPTICS 算法、DENCLUE 算法等[29-33]。

(4) 基于网格的方法

基于网格的方法(Grid-based Methods)首先将数据空间划分成为有限个单元(cell)的网格结构,所有的处理都是以单个单元为对象的。这样处理的优点就是处理速度很快,通常与目标数据库中记录的个数无关,它只与把数据空间分为多少个单元有关。代表算法有:STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法[7,34]。

(5) 基于模型的方法

基于模型的方法(Model-based Methods)是给每一个聚类假定一个模型,然后去寻找能够很好满足这个模型的数据集。这样一个模型可能是数据点在空间中的密度分布函数或者其他。它的一个潜在的假定就是:目标数据集是由一系列的概率分布所决定的。通常有两种尝试方向:统计的方案和神经网络的方案[34-38]。

聚类分析作为一种非常重要的数据分析技术,引起了国内外学者的广泛关注,在各方面都取得了较大进展,但其中提出的许多聚类算法及其改进算法还存在许多不足之处。现将其存在的问题总结如下:

(1) 自动确定聚类数目是评判聚类方法优劣的关键指标。针对 K-均值方法聚类数目难以设定的缺点,人们提出了不少有效的改进措施,常用的方法是不断增加聚类的数目,每一次都计算聚类的目标函数值,从而取得聚类效果最好的分类数目。这样大大增加了算法的计算复杂度,降低了聚类算法的效率。

(2) 人们利用各种优化算法和 K-均值算法相结合,以解决 K-均

值算法对初始聚类中心敏感及算法易陷入局部最优解等缺点，但同时考虑能够自动确定聚类数目的算法并不多见，且算法的效率也较低。

(3) 现有的许多聚类算法采用平方误差的总和作为准则函数，则使算法倾向于凸形状、大小相似的簇，不能发现任意形状任意大小的簇。现在提出的大部分准则函数通常假定簇为某一个几何形状，如果在同样的数据集中存在不同的结构，那这些方法通常就不能适应了。

(4) 孤立点和噪声点等各个方面的因素都可能对聚类结果造成较大的影响，因此现有聚类算法仍存在较大的局限性。

除此之外，由于聚类技术应用的广泛和深入，许多不同应用领域对于聚类技术在其他方面提出了各自的要求，如处理不同类型属性的能力、处理高维稀疏数据的能力、对于确定输入参数的领域知识需求最小化等。当前学术界对于聚类算法所应追求的各方面性能大致总结如下：算法对于高维和(或)大量数据的可伸缩性、处理不同类型属性的能力、发现任意形状的聚类、对于决定输入参数的领域知识需求最小化、处理带噪声数据的能力、增量聚类和对输入次序的不敏感、基于约束的聚类、聚类结果的可解释性和可用性。

1.3 本书的主要研究内容

本书针对复杂多源数据的知识获取与知识发现问题，分别从模式识别中聚类和分类的角度，在聚类分析方面，主要研究了基于序号编码的改进遗传模糊聚类算法和基于弧度点对称距离的自适应动态聚类算法；在分类问题方面，主要研究了基于自适应遗传算法的有趣分类规则知识发现和基于无参核学习的 Laplacian 正则化最小二乘分类。

(1) 基于序号编码的改进遗传模糊聚类算法

针对已有的基于遗传算法的 FCM 聚类算法计算复杂度高的缺

点，在已有遗传算法与 FCM 算法相结合的算法基础上对编码方式进行了改进，提出了基于序号编码的改进 GFCA 算法。用两组标注数据集对提出的算法进行了仿真实验，实验结果证明了该算法能在不损失聚类准确率的前提下减少计算复杂度。

(2) 基于弧度点对称距离的自适应动态聚类算法

自适应动态聚类算法结合遗传算法和聚类算法各自的优点，采用动态变长编码方式，实现自动演化聚类数目。在本算法中，利用弧度点对称距离取代欧几里得距离，将点分配到不同的聚类中，既能够发现凸面的又能够发现非凸面的聚类，并且无论聚类的大小和形状如何，只要聚类拥有对称特性，该算法都能适用。本算法充分发挥了遗传算法的全局寻优能力和一般聚类算法的局部搜索能力，可以更好地提高聚类质量。

(3) 基于自适应遗传算法的有趣分类规则知识发现

在数据挖掘中，数据分类是一个很重要的研究方面，然而现有的成熟分类算法一般只是从数据库中发现准确度很高的规则，而对发现有趣分类规则却论述不多，针对这一问题，提出了一种基于遗传算法挖掘有趣分类规则的方法。首先，通过属性信息增益和设置属性信息增益权值，以及规则的准确度来构造有趣分类规则的适应度函数，实现了对有趣规则的客观评价和主观评价的统一；其次，为了防止进化过程过早收敛或降低收敛速度，使用了自适应的遗传算法。最后使用 JBuilder 2006 平台给出了实验结果，从数据库中发现了有趣的分类规则，验证了算法的有效性。

(4) 基于无参核学习的 Laplacian 正则化最小二乘分类

针对 Laplacian 正则化最小二乘分类算法中人为指定核函数类型导致算法性能难以达到最优的问题，提出一种基于无参核学习(Non-Parametric Kernel Learning，NPKL)的 NPKL-LapRLSC 算法。NPKL 算法直接从数据中学习能更好地描绘数据相似性特征的半正定核矩阵，有效避免了针对特定问题选择和构造合适的核函数、根据实际样本数据确定核函数的参数等难题，提高了 LapRLSC 算法在实

际复杂应用中的适应能力。实验结果表明，选取合适的核函数对分类效果有很大影响，而本方法具有较高的识别率，且更为高效，更具有扩展性。

(5) 复杂网络的社区发现方法

社区发现是近年来社会网络中的研究热点。针对近邻传播算法的偏向参数影响聚类结果的类簇，提出了一种基于自适应近邻传播的社区发现算法，该算法将模块度引入到算法迭代过程中，作为评价指标引导算法向最优的聚类结果运行，该算法同 AP 算法一样不需要用户设定输入参数，不需要预先指定社区数量。通过引入标签传播算法的思想，提出了一种基于改进标签传播的社区发现算法，引入局部度中心的概念，首先找出网络中所有局部度中心节点，给它们以及其各自的邻居节点分配一个唯一的标签，然后进行异步标签更新。为了提高复杂网络中社团检测的性能，提出了一种基于结构相似度仿射传播的社团检测算法。该算法首先选取结构相似度作为节点之间的相似性度量，并采用了一种优化的方法来计算相似度。其次将计算得到的相似度矩阵作为输入，采用快速仿射传播算法进行聚类，得到最终的社团结构。实验结果表明改进的算法在绝大多数情况下都得到了优于原算法的结果，与其他经典社区发现算法相比，改进的社区发现算法也是有效的。

1.4 本书结构

本书一共 9 章，具体安排如下：

第 1 章是绪论部分。主要论述了课题的选题背景和研究意义，介绍了国内外复杂数据的知识获取与知识发现方法，重点讨论了数据的聚类分析和分类方法，并给出了本书的研究内容和结构。

第 2 章研究了基于序号编码的改进遗传模糊聚类算法。在传统遗传模糊聚类算法的基础上，提出了一种基于序号编码的改进GFCA算法，有效减少了遗传算法收敛所需的进化代数，减少了时间开销，

虽然在精度上有一些损失但已经可以为 FCM 提供合适的初始聚类中心来避开局部极值点，并进行了仿真实验验证。

第 3 章研究了基于弧度点对称距离的自适应动态聚类算法。在分析了遗传算法的基础上，提出了一种新的基于变长编码遗传算法的聚类分析方法。算法基于序号对聚类中心进行编码，编码的长度是在某一范围内随机产生，每一条染色体代表一种聚类结果；并给出了一个新的聚类指标作为适应度函数。最后，通过典型数据集验证了算法的有效性。

第 4 章研究了基于自适应遗传算法的有趣分类规则知识发现。本部分研究采用遗传算法挖掘有趣分类规则，提出了利用属性信息增益和设置属性信息增益权值来评价规则有趣度的方法，从而实现了规则有趣度的主观评价和客观评价的统一。

第 5 章研究了基于无参核学习的 Laplacian 正则化最小二乘分类。核函数是机器学习中核心组成部分，针对不同的分类问题，在使用时所采用的核函数及核函数参数选择会表现不同的分类特性。本部分研究以 LapRLSC 为学习模型，实现核函数的自动学习，提出一种基于无参核学习的 Laplacian 正则化最小二乘分类算法（NPKL-LapRLSC），与 LapRLSC 方法相比，NPKL-LapRLSC 算法具有较高的识别率，且更为高效、更具有扩展性。

第 6 章研究了基于自适应近邻传播的社区发现算法。针对近邻传播算法的偏向参数影响聚类结果的类簇，提出了一种基于自适应近邻传播的社区发现算法，该算法将模块度引入到算法迭代过程中，作为评价指标引导算法向最优的聚类结果运行，该算法同 AP 算法一样不需要用户设定输入参数，不需要预先指定社区数量。

第 7 章研究了基于改进标签传播的社区发现算法。通过引入标签传播算法的思想，提出了一种基于改进标签传播的社区发现算法，引入局部度中心的概念，首先找出网络中所有局部度中心节点，给它们以及其各自的邻居节点分配一个唯一的标签，然后进行异步标签更新。