基于分形维数的数据挖掘技术研究综述*)

倪丽萍　倪志伟　吴　昊　叶红云
(合肥工业大学管理学院　合肥 230009)

摘　要　分形维数在数据挖掘领域起着非常特殊的作用，它能有效地描述数据集，能反映复杂数据集中隐藏的规律性，基于分形维数的数据挖掘技术研究越来越受到人们的广泛关注。本文首先介绍了数据集的分形维数，进而在此基础上重点介绍了几种基于分形维数的数据挖掘技术，并对每种技术的特点进行了阐述，最后指出今后的发展方向。

关键词　分形，分形维数，数据挖掘

A Survey of the Research about Data Mining Technology Based on Fractal Dimension

NI Li-Ping　NI Zhi-Wei　WU Hao　YE Hong-Yun
(School of Management, Hefei University of Technology, Hefei 230009)

Abstract　Fractal dimension plays a very special role in data mining area. It can describe the data set effectively and can reflect the hidden regularity of the complex data set. Data mining technology based on fractal dimension has recently gained increasing attention all over the world. This paper first introduces the fractal dimension of the data set. Then focuses on some of the data mining technology based on fractal dimension and describes the characteristics of each technology. Finally points out the direction for the future development.

Keywords　Fractal, Fractal dimension, Data mining

1　引言

1989 年 8 月，第 11 届国际人工智能联合会议的专题研讨会上，首次提出了基于数据库的知识发现(KDD)技术。随后的 1995 年，美国计算机年会 ACM，又提出了数据挖掘(DM)的概念。由于数据挖掘是知识发现的一个重要步骤，因而往往将两者视为同一个概念。所谓数据挖掘技术，是指从大量数据中提取或“挖掘”知识[1]，从而帮助决策者做出正确的判断。数据挖掘技术自提出以来，得到了快速发展，并广泛应用于金融业、电信业、零售业等，然而，就数据挖掘技术的研究现状来看，目前仍然存在着一些问题，特别是如何更好地在复杂以及高维数据集上进行数据挖掘。

分形理论是现代非线性科学研究中十分活跃的一个数学分支，它的基本思想是利用整体与局部相似的特点，将一个复杂现象看成是由简单现象迭代而成，从而揭示复杂现象中所蕴含的规律和特性，特别适合于解决复杂问题。对于具有分形特征的物体而言，分形维数是一个重要的指标，它能够定量地描述分形集的复杂程度。近几年来，有研究表明分形维数在数据挖掘领域有着非常特殊的作用，将分形技术应用于数据挖掘领域能够更好地克服传统数据挖掘技术的不足，更加有效地解决在结构复杂、高维数据集上的数据挖掘问题。本文首先介绍几种基于分形维数的数据挖掘技术，进而探讨其发展方向。

2　数据集的分形维数

随着数据挖掘技术的发展，复杂数据挖掘成为我们现在所研究的重点，即如何在一些结构复杂、大数据集，例如，生物数据集、网络数据集、文本数据集、时序数据集上发现有用的知识。这些数据集结构复杂，维数较高，有的甚至随时间动态变化，因而给数据挖掘带来了新的挑战。事实上，这些数据集往往具有分形特征，即数据集的部分分布有着与整体分布相似的结构或属性[2]，可以利用分形理论进行分析。

分形维数是对分形集自相似性进行数学描述的一个方法，针对不同的研究对象，分形维数计算的方法也有所不同，因而分形维数有多种类型和定义方式，如：豪斯道夫维数、信息维、关联维、容量维等。一个自相似的数据集，分形维数体现了整个数据集的固有特性。为了说明分形维数在数据集中的作用，首先介绍两个与数据集有关的维度概念，一个是嵌入维，一个是本质维。其中嵌入维指该数据集所嵌入的维数，即数据集的整个属性集 F。本质维是描述一个数据集所需要的真正维数，即能够保持数据本质特征的属性集，小于或等于嵌入维。一般地，一个数据集的嵌入维并不能反映出数据集的真正特性，它无法表示数据集中是否存在相互关联的属性，以及属性的相关性情况。当数据集中存在相关属性后，数据集的本质维数就会相应地减小[3]，从而小于嵌入维，因而本质维描述了一个数据集真正需要的属性数。由于分形维数反映的是整个数据集的固有特性，因此通常用分形维数来表示一个数据集的本质维数。

数据集的分形维数可以采用如下的方法进行计算[4]，设一个 n 维的数据集，数据点落在边长为 r 的 n 维单元格内的概率为 p_i，则分形维数为：

$$D_q=\frac{1}{q-1}\frac{\log\sum p_i^q}{\log r} \tag{1}$$

由于数据集中的数据点是有限的，因此其仅在一定的区

*)本文受国家自然科学基金重点项目(70631003)资助。倪丽萍　博士生，助理研究员，研究方向为数据挖掘，机器学习；倪志伟　教授，博导，从事人工智能、数据挖掘、机器学习等方面的研究。

间尺度内具有统计自相似性，即 $r\in[r_{min},r_{max}]$。在(1)式中当 $q=0$ 时为豪斯道夫维数，$q\to1$ 时为信息维，而 $q=2$ 时为关联维。研究表明[4]，在这些分形维数中，信息维和关联维对于数据挖掘特别有用，其中信息维的变化说明了数据集变化的趋势，关联维数的变化则表明数据集中数据点分布的变化情况。因此利用分形维数的变化情况可以挖掘出数据集中的一些有用信息。目前，将分形技术应用于数据挖掘领域都是利用分形维数进行的。

3 基于分形维数的数据挖掘技术

3.1 基于分形维数的特征属性选择方法

特征属性选择通常是数据挖掘的一个前期工作，在数据挖掘中起着重要作用。属性选择的目的在于针对特定的应用选择最小的属性子集，且该子集能够保证不丢失数据的原有价值。通过特征属性的选择，能够提高数据的质量，加速挖掘速度。特征属性选择方法种类很多，可以从搜索方向、搜索策略、评价准则和停止标准四个方面来分析属性选择方法[5]。

基于分形维数的特征属性选择方法最早是由 Traina 在 2000 年提出的[6]。该方法也是目前基于分形维数属性选择方法中一个最具有代表性的算法。该算法在搜索方向上属于后向搜索，即首先考虑整个特征集 F，然后依据某种评价准则不断从 F 中选择最不重要的属性，直到达到某种停止标准。搜索策略采用启发式的搜索方式。而评价准则和停止标准是该方法的特别之处，主要利用上节中提到的数据集的分形维数来进行判断。

事实上，一个数据集的分形维数是不会随着冗余属性的加入而发生显著变化的，因而可以将分形维数的变化情况作为评价标准。即在后向搜索的每一步中，执行如下的操作：

Step1. 从当前属性集中，依次去除一个属性 i，并计算去除属性后数据集的部分分形维数 pD_i，其中 pD_i 表示去除属性 i 后的数据集的分形维数。

Step2. 选择部分分形维数与当前分形维数变化最小的属性 i，该属性为当前最不重要的属性，可以去除，去除该属性后将 pD_i 设置为当前分形维数。

上述步骤执行的终止条件为，剩余的属性个数为分形维数的上界，即当剩下的关键属性子集个数为原始数据集分形维数的上界时，终止后向搜索过程，所得到的为要选择的属性子集。

基于分形维数的属性选择方法特点在于利用分形维数的变化情况作为衡量属性重要性的一个标准。这种方法不需要对属性进行旋转和变换，因此能够很容易解释最终所得到的属性集，而且能对属性的重要性进行排序[6]。除此之外，分形维数说明了需要保留的关键属性的个数，在与其他属性选择方法结合时可以作为属性选择的一个停止标准。

由于该方法在计算分形维数时需要多次扫描数据集，因此一些研究者对该方法进行了改进。改进的目标在于减少扫描数据集的次数，提高计算分形维数的效率。例如，文[2]提出了一种基于分形维的快速属性选择算法(IFAS)，通过动态调整分形树(FD-tree)计算分形维数，只需扫描数据集一次。文[7]提出了一种基于 Z-ordering 的属性约简方法(ZBFDR)，该方法也只需要扫描数据集一次。文[8]将遗传算法与 ZBFDR 算法结合形成 GAZBFDR 算法，利用遗传算法的特点进一步加速了属性选择的速度。

3.2 基于分形维数的聚类方法

聚类是数据挖掘领域常用的一种方法，基于分形维数的聚类算法(FC)主要思想是利用同一簇的数据点之间的自仿射性或自相似性比不同簇数据点间的自仿射性或自相似性强的特点，根据维数的变化情况进行聚类。

该方法首先对数据集进行初始化聚类，设初始化后得到 n 个初始聚类，分别记为 $c_1,c_2,\cdots,c_n$。然后进行增量聚类，即设置阈值 δ，依次将每一个点加入到各簇 $c_1,c_2,\cdots,c_n$ 中，如果各簇分形维数的变化范围在 δ 内，则将该点加入维数变化最小的一簇 c_i 中，否则将该点视为离群点，单独作为一簇。具体描述如下[9]：

1)初始化阶段

设初始化后共聚成 k 个类，记为 $\{C_1,C_2,C_3,\cdots,C_k\}$。分别对初始化的类计算其分形维数，第 i 个类的分形维记为 $F_d(C_i)$。

2)增量聚类阶段

Step1. 从内存中调入一组数据集 $\{S\}$

Step2. 对于每一个点 $p\in S$

Step3. 对每一个类 $i=1,\cdots,k$

Step4. $C_i'=C_i\cup\{p\}$

Step5. 计算 C_i' 的分形维数

Step6. 在 k 个类中找到该点加入后使得原来类的维数变化最小的一个类，记为 $\hat{i}$，$\hat{i}=\min_i(|F_d(C_i')-F_d(C_i)|)$

Step7. 如果 $|F_d(C_{\hat{i}}')-F_d(C)_{\hat{i}}|$ 小于预定的一个阈值 ε，则认为该点属于类 $\hat{i}$，否则，认为该点为噪声数据。

其中，初始化阶段可以利用原来已有的聚类算法进行，例如 k-means 方法、基于距离的方法等。由上述步骤可以看出，基于分形维数的聚类算法认为不同簇的分形维数存在着较大的差异。

该算法相对于传统聚类算法来说，具有很多优势。首先，它是利用相似性进行聚类的，而不是利用基于距离的方法，因此能够更好地将不同类的数据区分开。其次该方法能够发现任意形状的聚类[9]，便于人们解释聚类的结果。

目前研究者在一些数据集上如：Web 负载记录[10]、历史气象数据集[11]测试基于分形维数聚类算法的有效性，都取得了较好的效果。由于该算法采用的是一种动态的增量方法，因此特别适合应用于数据流挖掘中，即适合对一些随时间呈现变化趋势的数据集进行挖掘。除此之外，根据该算法的思想，还可以将其应用于离群数据挖掘中，设置一定的阈值，当数据点落入数据集中，数据集的分形维数发生的变化大于设定的阈值，则说明该数据点为离群数据。使用这种方法进行离群数据挖掘有利于我们解释离群点的含义。

3.3 基于分形维数的关联规则发现方法

关联规则是数据挖掘中的一项重要技术，用于发现大量数据中项集之间的有趣关系，所发现的关联规则可以成为其他领域中进行决策的依据。

目前，基于分形维数的关联规则发现方法主要应用于量化关联规则挖掘。所谓量化关联规则是指数据集合中的数据不仅仅包含二进制属性数据，而且包含数字属性的数据，例如年龄属性等。

最为典型的基于分形维数的关联规则挖掘算法是 Barbara 等人于 2000 年提出的，该算法依然是利用分形维数的变化情况来进行关联规则的挖掘。动态计算数据集的分形维数，如果增加一个项集后，数据集的分形维发生了急剧变化(超过一定的阈值)，那么就认为是数据区间的一个分割点。详细的

算法描述可以参见文[12]。

该算法相对于以往的量化关联规则挖掘算法来说，最大的优势就在于能够很好地发现可能丢失的频繁项集。

3.4 基于分形维数的分类与预测方法

从广义上说，分类是预测的一种，分类是预测离散或分类号的[1]。而预测不仅包括离散值的预测还包括连续值的预测。分类和预测应用于数据挖掘领域可以描述重要数据类的模型或预测未来的数据趋势。与前面提到的几种基于分形维数的数据挖掘技术不同的是，基于分形维数的分类和预测主要是将分形维数作为一个特征量，将其与其他的分类或预测技术相结合从而进行分类或预测。例如，Ralph Hippenstiel等人结合分形维与神经网络对数字信号进行了分类；安国成等人在探讨了图像分类中三种分形维数提取方法的基础上，将分形维数与模糊c均值算法相结合对SAR图像进行分类；陈辉等人将分形维数与神经网络相结合构建识别模型，并以地震检测波为分析对象，对混凝土地下连续墙的区域物性进行预测等。在这些应用中都是将分形维数作为描述事物的一个特征量的。

值得一提的是，利用分形维数进行预测时，除了使用上面提到的方法外，还有一个比较常用的方法，利用分形统计模型进行分析预测。一般地，分形统计模型可以表示如下[17]：

$$N(r)=Cr^{\pm D} \tag{2}$$

其中r表示特征尺度，$c>0$为比例常数，$D>0$称为分维数，$N(r)$表示尺度大于等于或小于r的个数，其中当分形维数前面取负号时，记为$N\geqslant r$，否则记为$N\leqslant r$。从该式可以看出分形分布的特点要求大于或小于某一尺度物体的数目，与物体大小之间存在幂函数的关系。事实上，在现实中有很多现象满足幂率分布，例如，人们的收入分布，网站的点击率，图书馆文献的引用率等[18]，因此都可以利用分形分布统计模型作为预测模型。在使用时，利用最小二乘法求出式(2)中的分形维数，进而得到分形统计模型的各参数，将其作为预测的数学模型，解释现实现象。目前该方法已广泛应用于地质以及股票领域，对地质矿产以及股票走势进行预测。

由于分形维数往往描述了物体的性质，因此将分形维数作为一个特征量，能够有效地提高分类和预测的准确率。

4 基于分形维数的数据挖掘技术进一步研究方向

基于分形维数的数据挖掘技术，虽然近几年来受到了人们的高度重视，但是由于它的研究难度较大，目前尚处于起步阶段，还有很多问题值得研究，例如，如何判断数据集具有分形特征，如何快速计算数据集的分形维数，如何在计算机上模拟实现，如何解释数据集分形维数的实际意义等。因而进一步的研究方向如下：

(1)与实际应用相结合。目前基于分形维数的数据挖掘技术研究还多处于实验阶段，将这种技术应用于实际是下一步需要考虑的问题。

(2)多重分形维数的研究。就目前基于分形维数的数据挖掘技术看，算法中所使用的分形维大多是豪斯道夫维、信息维、关联维数等一些简单分形维数，然而仅仅使用这种简单分形维数往往不够准确，因此下一步主要的研究目标是多重分形维数，从而更加准确地描述数据集的特性。

(3)与其他智能技术的结合。目前，基于分形维数的数据挖掘技术与其它智能技术的结合也有部分研究成果，例如，上述提到的与神经网络的结合，与遗传算法的结合等，但是，这种结合性的研究范围还比较局限，下一步可以考虑进一步扩大分形维数的应用领域，将其与更多的智能技术相结合，从而提高解决问题的能力和效率。例如，分形与CBR的结合，分形与其他聚类算法的结合等。

(4)算法的进一步改进。上述所提到的基本算法还存在一些不足之处，可以对其进行进一步的改进，例如，提高算法的效率，更加合理地设置算法中的阈值等。

参考文献

1 Han Jiawei, Kamber M. 数据挖掘概念与技术. 范明，孟小峰，等译. 北京：机械工业出版社，2001

2 鲍玉斌，王琢，孙焕良，于戈. 一种基于分形维的快速属性选择算法. 东北大学学报(自然科学版)，2003，24(6)：527～530

3 Lee H D, Monard M C, Wu Feng Chung. A Fractal Dimension Based Filter Algorithm to Select Features for Supervised Learning. IBERAMIA-SBIA, 2006. 278～288

4 Barbara D. Chaotic Mining: Knowledge discovery using the fractal dimension. In: 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Philadelphia USA, 1999

5 彭佳红，沈 岳，张林峰. 数据挖掘中的特征选择及其算法研究[J]. 计算机工程与设计，2005，26(5)

6 Traina Jr C, Traina A, Wu L, Faloutsos C. Fast feature selection using fractal dimension. In: Proceedings of XV Brazilian Symposium on Databases, Paraila: Springer, 2000. 78～90

7 Yan Guanghui, Li Zhanhuai, Yuan Liu. The Practical Method of Fractal Dimensionality Reduction Based on Z-Ordering Technique. ADMA, 2006. 542～549

8 Yan Guanghui, Li Zhanhuai, Yuan Liu. On Combining Fractal Dimension with GA for Feature Subset Selecting. MICAI, 2006. 543～553

9 Barbara D, Chen P. Using the fractal dimension to cluster datasets. Knowledge Discovery in Databases. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston USA, 2000. 260～264

10 Menasce D A, Abrah ao B D, Barbara D, Almeida V A F, Ribeiro F P. Fractal characterization of web workloads. In: Proceedings of the 11th International World Wide Web Conference, 2002

11 Barbará D, Chen Ping. Tracking Clusters in Evolving Data Sets. FLAIRS Conference, 2001. 239～243

12 Barbara D, Nazeri Z. Fractal Mining of Association Rules over Interval Data: [Technical Report]. George Mason University, 2000. 9

13 Georgieva T. Using the Fractal Dimension of Sets to Discover the Distribution Intervals of Association Rules in OLAP Data Cubes. In: Proceedings of the First International Conference on Information Systems and DataGrids, Sofia, 2005. 88～98

14 Hippenstiel R, El-Kishky H, Radev P. On Time-Series Analysis and Signal Classification-Part I: Fractal Dimensions

15 陈辉，李益进. 地震检测波的分形神经网络模式识[J]. 建筑技术开发，2005，32(12)：39～43

16 安国成，刘振华，于文震. 基于分形特征的高分辨率SAR图像分类[J]. 现代雷达，2006，28(6)：26～29

17 申维著. 分形混沌与矿产预测[M]. 北京：地质出版社，2002

18 Faloutsos C. Data Mining using Fractals and Power laws. http://www.cs.cmu.edu/～christos/TALKS/MSU-05/msu05-dl-v02.pdf. 2005. 10. 21