文章编号:1000 - 582X(2004)12 - 0111 - 05

小波在数据挖掘算法中的运用*

文俊浩[1],胡显芝[1],何光辉[2],徐 玲[1]
(1. 重庆大学 软件学院,重庆 400030;2. 重庆大学 数理学院,重庆 400030)

摘 要:由于小波理论具有良好的构造性和实际应用性,近年来被广泛地应用于诸如图像处理、计算机可视化、网络管理和数据挖掘等计算机科学研究领域。小波有很多良好的性质,如多分辨的分解结构、变换的时空线性复杂性等特性,从而可以为数据挖掘提供更加有效的算法。给出了小波在聚类、分类、分布式数据挖掘、相似性搜索、近似查询处理、可视化等算法中的运用,讨论了小波在数据挖掘研究中的影响,并简述了有潜力的未来研究方向。

关键词:数据挖掘;小波;数据挖掘算法
中图分类号:TP391 **文献标识码**:A

小波变换是多年来从数学和信号处理等许多不同领域形成的一种综合方法。一般来说,小波变换是一种将数据、函数或算子分解成不同频率的成分,然后在与其尺度相匹配的分辨中研究各个成分的工具。因此,小波变换就被用来对许多有兴趣的对象提供实用和有益的数学表达式。现在许多计算机软件包中包含有进行小波变换的快速有效的算法,这样就可以很容易地实现小波变换,因此,无论是在理论研究还是应用领域,小波变换受到研究人员普遍的欢迎。总之,小波被广泛地应用于诸如图像处理、计算机可视化、网络管理和数据挖掘等计算机科学研究领域。

在过去的十多年里,数据挖掘在学术研究和应用方面成了一个重要领域。数据挖掘是从大量数据中抽取挖掘出未知的、有价值的模式或规律等知识的复杂过程。小波理论具有良好的构造性和实际应用性,因而可以在数据挖掘过程中发挥重要的角色。小波有很多良好的性质,如消失矩、分级的和多分辨的分解结构、变换的时空线性复杂性、不相关系数和基函数的广泛多样性,这些性质可以运用于许多数据挖掘问题,从而为数据挖掘提供更加有效的解决方法。用小波描述数据挖掘的数据,可以使数据挖掘过程更加有效和准确,小波可以渗透到许多数据挖掘算法中。尽管标准的小波是应用在有时空局限性的数据上(如时间序列、流数据和图像数据),但很多小波的相关方法被应用在大范围的数据挖掘问题上。笔者总结和归纳了小波在数据挖掘上的应用,以引起数据挖掘研究者更广泛的注意。

1 数据挖掘的任务和算法

数据挖掘的任务和算法涉及到一些必要的步骤,在这些步骤中,用了许多巧妙的方法来提取有用的信息模式。数据挖掘的任务有多种,比如:聚类、分类、内容检索和可视化等,每种功能都可以看作是用数据挖掘算法解决的一个特殊问题。一般说来,许多不同的算法可以解决同一个任务。同时,某些算法也可以应用在不同的任务上。笔者将简述小波在数据挖掘和算法中的多种应用,根据任务的不同进行较系统的讨论,主要包括聚类、分类、分布式数据挖掘、相似性搜索、查询处理和可视化等。

1.1 聚 类

聚类问题产生于许多学科中,并且有着广泛的应用。聚类问题可以直观的按以下分类:假设 w 为多维空间中 n 个数据点的集合,找出一个 w 的划分,将划分成类,使得类中的点都彼此相似,类间的点尽可能不

* 收稿日期:2004 - 10 - 12
基金项目:重庆市自然科学基金资助项目(2004BB2182)
作者简介:文俊浩(1969 -),男,河南临颍人,重庆大学副教授,博士研究生,主要研究方向为数据挖掘、软件工程。

同。聚类问题可以从不同的角度和不同的方法进行研究,如机器学习、数据库和数理统计等。小波变换的多分辨分析的特性启发了研究者研究某些算法,使聚类算法具有多分辨率的特性。

对超大的空间数据库而言,小波聚类算法[1]是一个多分辨分析聚类。空间数据对象可以用 n 维特征空间表示,并且空间对象的数字特征也能用特征向量表示。在这些向量中,向量的每个组成部分对应于一个数字特性。用网格对数据空间进行分类不但可以减少数据对象的个数,并且误差也很小。从一个信号的处理角度来看,如果特征空间的对象集合被看作是 n 维信号,信号的高频部分对应于这样一个区域,在区域中,对象的分布有一个快速的变化(即聚类的边缘),而 n 维信号的低频部分对应着特征空间的高幅度的区域,在这个区域中,对象是被集中的(即聚类),用小波变换将信号分解为不同频率的子带。因此,为了确定聚类,也就转换到与特征空间相联系的组成部分。而且将小波应用于特征空间能提供多分辨分析的数据描述,所以寻找不同的相互联系的组成部分可以在不同的分辨尺度下进行。换句话说,小波变换的多分辨特性能使小波聚类算法在不同的尺度下以任意的精确度有效地确定任意形状的聚类。实验表明,小波聚类算法效率大大优于 Brich[2] 和 CLARAS[3],并且它是一种稳定和有效的聚类方法。Sheiknoleslami[4] 提出了一种基于内容的搜索方法,这种方法利用小波变换提取几何图像的不同数字特征。利用基于小波的多尺度分解,通过聚类方法两类不同性质集合都可以系统地表示出来。对每一个特征集合,在数据库中对聚类图像设计和实现不同距离的测量技术。实验表明,当运用基于小波的聚类方法时,可以提高搜索的效率和效果。

1.2 分 类

数据分类是为了确定可以表明每个实体所属的群的特性。分类既可以用来理解已存在的数据,也可以预测新实体如何工作。数据分类包含了许多算法,如判定树归纳、贝叶斯分类和贝叶斯网络、神经网络,K-最邻近分类、基于案例的推理、遗传算法、粗糙集和模糊逻辑技术。针对较大的二维数据集合,特别是较大的数字图像的分类问题,Csatelli 等人提出了基于小波的分类算法,小波对于分类的实现是非常有用的。首先,可以将分类方法应用到原始数据的小波域。其次,可以将小波的多分辨分析的特性加入分类过程,使得过程更简洁。该算法是把图像看作实值二维数组,二维数组中的每个元素是一个像素,并且每个像素与一个点向量相联系,一个标签表示它的类。分类问题主要由两大类组成:一类是用已知点但未知标签观察图像;另一类问题是为每个点分配一个标签。在数字图像中,需要快速和有效地分类大的图像,第 2 类问题主要是由这个需求所激发的。传统的方法主要是逐点分析法,这种方法除了计算量特别大外,也没考虑到相邻点之间标签的正确性。基于小波的分析方法是一种建立在高级的分类框架基础上的方法,它的核心思想如下:在数据的低分辨描述上运用一般的分类方法,而这些数据是由离散的小波变换得到的。小波变换产生了数据的多分辨格式描述。在这个描述中,每个尺度的一个点都对应原始图像中 $k \times k$ 阶点组成的片。在分类的每一步过程中,算法决定每个系数是否对应类似的点片并且将相同的标签分配给整个片或在更高尺度下分配相同标签以重新检查数据,并且重复此过程直到获得满意结果。基于小波的分类方法与传统的点分类方法相比,在速度上有一个重大的突破。对高度相关的点值图像而言,这种方法将比相对应的非先进的分类法得到的结果要精确得多,因为离散小波产生对应于 $k \times k$ 片的权值平均,并且,这种算法与看单个图像时可能出现的相比,具有更优的一致性。Castalli 等人在卫星图像的分类中运用了这种方法,并且给出了这种方法的原理分析。Blume 和 Ballard[5] 基于学习向量量化和局部化的 Harr 小波变换特征,描述了一种将图像点分类的方法。Harr 小波变换被用来在各个图像点产生一个特征向量,并且它提供了和周围区域结构一样的局部亮度和色彩信息。用手标签的图像被用作产生电报密码本,这种密码本是利用最佳学习效率学习向量量化的算法。对于少量类,点分类高达 99%。Scheunder 等人[6] 的基于小波变换详细拟定了研究内容,使多分辨和正交描述在结构分类和图像分割中扮演重要角色。使用离散小波变换可以提取有用的灰度级和色彩结构特性,发现有用的旋转特性。Mojsilovic 等人[7] 提出了用基于小波的小维度的方法来解决结构抽样的分类。Tzanetakis 等人在文献[8]中用小波提取特征集合来描述音乐表面和节奏信息,从而给出了建立自动类型的分类算法。

1.3 分布式数据挖掘

近年来,随着技术的进步,计算能力以及存储能力的日益提高,互联网对日常生活的渗透和商业、制造业、科学研究的日趋自动化,数据集规模有了迅速的增长。而且这些数据集的大部分都按地理位置分布于多个场所。为了挖掘如此巨大且分布式排列的数据集,研究出一种有效的分布式算法将是非常重要的。这样不但可以降低通信成本,而且可以减少中心存储需求

以及算法时间复杂度。在一个分布式的环境中，数据集可能是同类的，例如，不同站点包含了具有完全相同特征集；也可能是异类的，比如说，不同的站点存储了不同特征集，在不同的站点之间也有可能具有普遍特征。小波基的正交特性在分布式数据挖掘中扮演一个重要的角色，这是因为正交性可以保证正确地和独立地进行局部化分析，并且这种分析可以作为整体模型的模块。此外，由于小波具有紧支性，所以利用小波处理本地数据的同时不会影响其他区域的数据，从而小波可以用来设计并行算法。

Kargupta 等人[9]提出了从不同种类站点进行分布式数据挖掘的思想，该思想使用了基于小波的整体数据挖掘(CDM)方法，主要步骤如下：

1）选择一个正交表示，该正交表示适合于将被建构数据模型的类型；

2）在每个局部站点里产生一个近似的标准正交基系数；

3）从每个站到单一站，移动数据集的近似选择样本，并生成对应的非线形交叉项的近似基系数；

4）合成局部模型，将模型转化成用户可以理解的形式，并输出模型。

CDM 的原理基于这样一个事实：如果使用的基函数适当，任何函数可以用分布的形式表示出来。如果使用小波基，根据小波基的正交性，可以保证正确地和独立地进行局部化分析，并且这种分析可以作为整体模型的模块。Hershberger 等人[10]把这种基于小波 CDM 方法用于多元回归分析和线性判别分析。实验证明用 CDM 所得到的结果优于用集中式方法得到的结果。

1.4　相似性搜索

在数据挖掘中相似搜索的是：给出感兴趣的模式，在数据集中基于相似测度找出相似模式。小波将几种不同的方法应用于相似搜索。首先，小波将原始数据变换到小波域，可以仅依靠小波系数去实现维数的降低，相似搜索就可以在小波域中进行，且效率较高。值得一提的是对于这里的数据集，用小波将 n 维空间投射到 k 维空间，同样的 k 小波系数将保存在数据集里，很明显对于所有的目标来说这并不是最理想的，为了在数据集中找到 k 个优化系数，需要计算每个系数的平均权重；第二，小波变换被用于求解紧特征向量，并定义新的相似测度以加快搜索；第三，小波变换能在不同的尺度下支持相似搜索，而这种相似程度又是用可适应的交互式的方法来定义。

小波已经广泛地应用于时序上的相似搜索[11]。Chan 和 Fu 提出了小波中的有效时序匹配方法。Huhtala 等人在连续的时序中使用小波通过寻找相似点来找出共同特征。Wu 等人给出在时序匹配上 DFT 与 DWT 不同点。实验结果表明，DWT 虽然不能降低相关匹配错误，也不能提高在相似搜索中查询的精确程度；然而，DWT 有几个优点，比如 DWT 有多分辨分析特征，DFTD 的复杂程度是 $O(N\log N)$，而 DWT 的复杂程度是 $O(N)$。小波变换给出了信号的时频局部化。所以信号能量的大部分仅依靠一些 DWT 系数就可以表达出来。Struzik 和 Siebes 给出了一种新的相似测度，它利用基于 Haar 小波变换的特殊方法取代了保持小波系数的选择性方法。

基于内容的小波相似性搜索方法被广泛地应用于图像与音频数据库。Jacobs[12]等人提出了一种基于内容的使用图像查询距离快速而有效的图像查询方法，图像查询距离是由通过剪枝和量化的小波分解而获得的小波特征标识来计算的。Natsev 等人提出了用于相似搜索的 WALRUS(基于小波的用户指定场景搜索)算法，Ardizzoni[13]等人提出了一种 WINDSURF(基于小波的使用区域划分的图像搜索)图像搜索的新方法。WINDSURF 首先使用 Haar 小波变换提取图像的颜色和纹理特征，接着使用聚类技术把图像分割成区域，然后计算相似性，并把它作为匹配区域之间的 Bhattcharyya[14] 距离。Wang[15-16] 提出了一种新的 WBIIS(基于小波的图像索引和搜索)图像索引和搜索算法，它对于大量的图像数据库有着部分轮廓搜索的能力。Wang, Wiederhold 和 Firschein[17] 提出了 WIPETM(基于小波色情图像过滤)图像搜索方法。WIPETM 使用了 DB-3 小波，归一化中心矩和彩色直方图，从而为相似匹配提供特征向量。Subramanya 和 Youssef[18]提出一种可变的基于内容的图像索引和搜索系统，该系统是基于彩色图像的高度不相关小波系位平面，这个位平面被用于搜索有效特征向量空间时，引出了一个基于内容的可升级的图像索引和源于色彩图像的向量系数的搜索方法。Mandal[19] 等人为图像索引提供了一个快速的小波直方图技术。

1.5　近似查询处理

由于许多决策支持应用具有试探特性，所以用户在大多数情况下不需要完全准确的查询响应，而更希望快速近似地查询响应。小波变换可用于数据压缩，所以通过小波变换可以把数据变为紧凑小波大纲的形式，然后再对小波大纲进行近似查询。小波大纲能把大量数据缩减为紧凑数据集，这就为查询提供了一种快速而合理的近似查询方法。

Matias, Vitter 和 Wang[20]提出了基于小波的直方图技术,该技术通过建立用于选择性估计量的潜在分布式数据的直方图来实现,而 Vitter[21]等人对于全局查询的近似性提出了一种基于小波的查询技术。这种技术的主要思想是对输入数据做多维小波分解,并通过保留选择性的小波系数的集合,来获得一个紧的数据大纲。文献[20]的实验表明了基于小波的直方图技术有利于提高查询精度,文献[21]也证明了该技术的有效性,并且可以减小构造成本和存储消耗。Chakrabarti 等人将前面的小波技术推广到了近似查询响应,对于决策支持,并证明了该小波技术是一种普通而有效的工具。其方法通常由以下 3 步完成:首先,计算小波系数大纲,然后用于查询语言 SQL 中,比如,选择、投影、合并都能在小波域里完全地执行,最后把结论从小波域映射到相关的表示域(Rendering)。

1.6 可视化

可视化是数据挖掘中形象描述数据的一种有效方法,由于图形比文本和数字隐含更多信息量,从而使得用户能够对数据形成一种直观理解。然而,对于大量的数据,执行简单的可视化过程是不可能的。多尺度小波变换对数据的观察,可以实现先观察数据最重要的特征,然后再渐进地观察数据的细节。

Miller 等人提出了一种基于小波的新方法,该方法主要是对松散的文本数据进行可视化研究,其主要技术是首先将文档的文字构建成传统的数字信号,然后对该信号做小波变换,最后在频域中运用多分辨分析能量对文档的叙述流的特征进行分析。Wang 和 Bergeron 讨论了数据分解的可靠性问题,详细分析了数据的可视化问题。对该问题的研究使用了 6 个数据集,其不但说明了紧支正交小波的特征,而且还提出了一种误差跟踪机制,该机制借助可利用的小波资源来测量小波逼近度的质量。Roerdink 和 Westenbery 认为运用小波可对大量数据集进行多分辨可视化。首先从数据的小波分解开始,接着计算低分辨率图像,根据小波的多分辨的特性,这样的近似度可以不断地提高。Du 和 Moorhead 提出了一种使用小波变换和 MPI(消息传送接口)来实现分布式可视化系统。实践证明小波变换在数据分解和渐进传输中是一个非常有用的工具。

2 结 论

小波在数据挖掘算法中的应用,主要是使用正交小波基。然而,虽然小波消失矩能实现去噪和降维的目的,但对于处理噪声数据来说不是最好的。直观地看,正交性是非常经济的方法。也就是说,在每个方面,它都能包含平等的重要信息。可是,当试图去除噪声或者不一致信息的时候(噪声也作为不一致信息的一部分),通常有可能小波系数的阈值也会使一些有用的信息丢失。为表示不一致信息,使用冗余小波的表示——小波框架也许会更好。除了不具有正交性之外,小波框架保留了其他标准正交小波基所拥有的其他所有特性,比如消失矩,紧支性,多分辨率。小波框架的冗余意味着框架函数不再独立了,比如向量(0,1)和(1,0)是 R^2 的标准正交基,但向量(1/2,1/2),(-1/2,1/2)和(0,1)却并不正交,不是 R^2 的基向量。因此,当数据中含有噪声时,小波框架将提供一些记录噪声的特别方向。将来的工作就是建立关于噪声方向和冗余信息的标准。

小波可能被用于其他一些新的研究与应用中,比如数据库压缩、多分辨的数据分析和快速近似数据挖掘等等。小波方法将在数据挖掘中有更长远的发展和应用。

参考文献:

[1] SHEIKHOLESLAMI G, CHATTERJEE S, ZHANG A. WaveCluster: A Multi-resolution Clustering Approach for Very Large Spatial Databases[A]. In Proc 24th Int Conf Very Large Data Bases[C]. New York: [s. n.], 1998. 428-439.

[2] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an Efficient Data Clustering Method for Very Large Databases [A]. In Proceedings of ACM SIGMOD[C]. Montreal: ACM Press, 1996. 103-114.

[3] NG R T, HAN J. Efficient and Effective Clustering Methods for Spatial Data Mining[A]. 20th International Conference on Very Large Data Bases[C]. Santiago: [s. n.], 1994 (12-15). 144-155.

[4] SHEIKHOLESLAMI G, ZHANG A, BIAN L. A Multiresolution Content-based Retrieval Approach for Geographic Images[J]. GeoInformatica, 1999, 3(2): 109-139.

[5] BLUME M, BALLARD D. Image Annotation Based on Learning Vector Quantization and Localized Haar Wavelet Transform Features [EB/OL]. http://www.reticular.com/Library/ImageAnnot.pdf, 2003-12-08.

[6] SCHEUNDERS P, LIVENS S, WOUWER G. Wavelet-based Texture Analysis[J]. International Journal on Computer Science and Information Management, 1998, 1(2): 22-34.

[7] MOJSILOVIC A, POPOVIC M V. Wavelet Image Extension for Analysis and Classification of Infarcted Myocardial Tissue [J]. IEEE Transactions on Biomedical Engineering, 1997,

44(9):856 - 865.

[8] TZANETAKIS G, ESSL G, COOK P. Automatic Musical Genre Classification of Audio Signals [EB/OL]. http://ismir2001.ismir.net/pdf/tzanetakis.pdf,2004 - 01 - 12.

[9] HERSHBERGER D E, KARGUPTA H. Distributed Multivariate Regression Using Wavelet-based Collective Data Mining [J]. Journal of Parallel and Distributed Computing, 2001, 61(3):372 - 400.

[10] HUHTALA Y, KARKKAINEN J, TOIVONEN H. Mining for Similarities in Aligned Time Series Using Wavelets [A]. In Data Mining and Knowledge Discovery: Theory, Tools, and Technology[C]. [s.l.]:SPIE Proc, 1999.

[11] JACOBS C E, FINKELSTEIN A, SALESIN D H. Fast Multiresolution Image Querying[J]. Computer Graphics, 1995, 29:277 - 286.

[12] ARDIZZON S, BARTOLINI I, PATELLA M. Windsurf: Regionbased Image Retrieval Using Wavelets[J]. Database and Expert Systems Applications, 1999, (1 - 3):167 - 173.

[13] BRITO M, CHAVEZ E, QUIROZ A, et al. Connectivity of the Mutual K-Nearest-Neighbor Graph for Clustering and Outlier Detection[J]. Statistics and Probability Letters, 1997, 35:33 - 42.

[14] WANG O F J Z, WIEDERHOLD G, WEI S X. Wavelet-based Image Indexing Techniques with Partial Sketch Retrieval Capability[J]. IEEE Advances in Digital Libraries, 1997, 1(4):311 - 328.

[15] WANG J Z, WIEDERHOLD G, FIRSCHEIN O. Content-based Image Indexing and Searching Using Daubechies' Wavelets[J]. International Journal on Digital Libraries. 1997, 1(4):311 - 328.

[16] WANG J Z, WIEDERHOLD G, FIRSCHein O. System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms [EB/OL]. http://www - db.stanford.edu/ ~ wangz/project/imscreen/WIPE/wang.pdf, 2004 - 03 - 05.

[17] SUBRAMANYA S R, YOUSSEF A. Wavelet-based Indexing of Audio Data in Audio/multimedia Databases [EB/OL]. http://ieeexplore.ieee.org/iel4/5744/15353/00709492.pdf, 2003 - 11 - 12.

[18] MANDAL M K, ABOULNASR T, PANCHANATHAN S. Fast Wavelet Histogram Techniques for Image Indexing [J]. Computer Vision and Image Understanding, 1999, 75(1 - 2):99 - 110,

[19] MATIAS Y, VITTER J S, WANG M. Wavelet-based Histograms for Selectivity Estimation [A]. In ACM SIGMOD [C]. [s.l.]:ACM Press, 1998. 448 - 459.

[20] VITTER J S, WANG M. Approximate Computation of Multidimensional Aggregates of Sparse Data Using Wavelets [EB/OL]. http://www.math.tau.ac.il/ ~ matias/courses/sem_fall98/wavelets.htmlj, 2004 - 02 - 15.

[21] CHAKRABARTI K, GAROFALAKIS M, SHIM K. Approximate Query Processing Using Wavelets[J]. VLDB Journal: Very Large Data Bases, 2001, 10(2 - 3):199 - 223.

Wavelet Application in Algorithms of Data Mining

WEN Jun-hao[1], *HU Xian-zhi*[1], *HE Guang-hui*[2], *XU Ling*[1]

(1. College of Software Engineering, Chongqing University, Chongqing 400030, China;
2. College of Science, Chongqing University, Chongqing 400030, China)

Abstract: Wavelets have been widely applied in such research areas as image processing, computer vision, network management and data mining. Wavelets have many favorable properties, which are hierarchical and multiresolution decomposition structure, linear time and space complexity of the transformations. The paper presents wavelet application on the algorithms of data mining, in cluding as clustering, classification, distributed data mining, similarity search, approximate query processing, visualization and so on. Finally, the paper concludes by discussing the impact of wavelets on data mining research and outlining potential future research directions and applications.

Key words: data ming; wavelet; algorithm of data mining

(编辑　张　苹)