数据挖掘中的软计算方法及其应用

Soft Computing And It's Applacation in Data Mining

汤效琴 （宁夏大学 数学计算机学院 宁夏银川市 750021）

摘 要： 文章对数据挖掘中软计算方法及应用作了一个综合性阐述。对模糊逻辑、遗传算法、神经网络、粗集等软计算方法，以及它们的混合算法的特点进行了分析，并对它们在数据挖掘中的应用进行了分类。

关键词： 数据挖掘 软计算 模糊逻辑 遗传算法 神经网络 粗集

1 引言

在过去的数十年中，随着计算机软件和硬件的发展，我们产生和收集数据的能力已经迅速提高。许多领域的大量数据集中或分布的存储在数据库中，这些领域包括商业、金融投资业、生产制造业、医疗卫生、科学研究，以及全球信息系统的万维网。数据存储量的增长速度是惊人的。大量的、未加工的数据很难直接产生效益。这些数据的真正价值在于从中找出有用的信息以供决策支持。在许多领域，数据分析都采用传统的手工处理方法。一些分析软件在统计技术的帮助下可将数据汇总，并生成报表。随着数据量和多维数据的进一步增加，高达 109 的数据库和 103 的多维数据库已越来越普遍。没有强有力的工具，理解它们已经远远超出了人的能力。所有这些显示我们需要智能的数据分析工具，从大量的数据中发现有用的知识。数据挖掘技术应运而生。

数据挖掘就是指从数据库中发现知识的过程。包括存储和处理数据，选择处理大量数据集的算法、解释结果、使结果可视化。整个过程中支持人机交互的模式[1]。数据挖掘从许多交叉学科中等到发展，并有很好的前景。这些学科包括数据库技术、机器学习、人工智能、模式识别、统计学、模糊推理、专家系统、数据可视化、空间数据分析和高性能计算等。数据挖掘综合以上领域的理论、算法和方法，已成功应用在超市、金融、银行、生产企业和电信，并有很好的表现。

软计算是能够处理现实环境中一种或多种复杂信息的方法集合。软计算的指导原则是开发利用那些不精确性、不确定性和部分真实数据的容忍技术，以获得易处理、鲁棒性好、低求解成本和更好地与实际融合的性能。通常，软计算试图寻找对精确的或不精确表述问题的近似解[2]。它是创建计算智能系统的有效工具。软计算包括模糊集、神经网络、遗传算法和粗集理论。

2 数据挖掘中的软计算方法

目前，已有多种软计算方法被应用于数据挖掘系统中，来处理一些具有挑战性的问题。软计算方法主要包括模糊逻辑、神经网络、遗传算法和粗糙集等。这些方法各具优势，它们是互补的而非竞争的，与传统的数据分析技术相比，它能使系统更加智能化，有更好的可理解性，且成本更低。下面主要对各种软计算方法及其混合算法做系统性的阐述，并着重强调它们在数据挖掘中的应用情况。

2.1 模糊逻辑

模糊逻辑是 1965 年由 Zadeh 引入的，它为处理不确定和不精确的问题提供了一种数学工具。模糊逻辑是最早、应用最广泛的软计算方法，模糊集技术在数据挖掘领域也占有重要地位。从数据库中挖掘知识主要考虑的是发现有兴趣的模式并以简洁、可理解的方式描述出来。模糊集可以对系统中的数据进行约简和过滤，提供了在高抽象层处理的便利。同时，数据挖掘中的数据分析经常面对多种类型的数据，即符号数据和数字数据。Nauck[3] 研究了新的算法，可以从同时包含符号数据和数字数据中生成混合模糊规则。数据挖掘中模糊逻辑主要应用于以下几个方面：

(1) 聚类：将物理或抽象对象的集合分组成为由

类似的对象组成的多个类的过程被称为聚类。聚类分析是一种重要的人类行为,通过聚类,人能够识别密集的和稀疏的区域,因而发现全局的分布模式,以及数据属性之间有趣的关系。模糊集有很强的搜索能力,它对发现的结构感兴趣,这会帮助发现定性或半定性数据的依赖度。在数据挖掘中,这种能力可以帮助阻止搜到无用和微不足道的知识。研究者为此发展了模糊聚类算法,并得到了广泛应用[4][5]。在高维数据挖掘中有太多的属性要考虑,因此知识简约就非常的必要。属性聚类的实质就是知识简约,所谓知识约简,就是在保持知识库的分类或者决策能力不变的条件下,删除不重要的或冗余的知识,最小约简(含有最小属性)是人们所期望的,且约简结果是不确定的。所以模糊聚类成为知识简约的有力工具。

(2) 关联规则:数据挖掘重要的一点是关联规则的发现,关联规则挖掘是寻找给定数据集中属性间的关联。其中,布尔关联规则考虑的是关联的属性在与不在的二维特征,概化关联规则描述的是属性的分层关系,量化关联规则描述的是量化的属性(既离散化的属性)间的关联[6]。由于使用模糊概念表示的规则更符合人的思维和表达习惯,增强了规则的可理解性,所以模糊技术已成为数据挖掘系统中的关键技术。文献[7]中用模糊分类开拓了概化关联规则。

(3) 数据概化:概化发现是数据挖掘重要部分之一。它将大的数据集从较低的概念层抽象到较高的概念层,用可理解的信息来表达数据库中最重要的部分,并提供给用户。

大数据集的语言概化通过有效的程度来获得,参考的标准内容在挖掘任务中。系统由概述、一致性程度真实和有效性组成。已经发现的最有兴趣的语言概化并不琐碎,却很人性化。实际上,它并不能自动地进行概化,需要人的操作。Kacprzyk 和 Zadrozny[8] 发展了功能依赖度,语言概化使用了自然和可理解性的词汇,它支持模糊元素,包括属性间模糊的、重要的相互作用。首先,用户必须制定概化兴趣度,然后系统从数据库中获得记录,并计算每个概化的有效性,最后,选择最适合的语言概化。此方法通过网络浏览器已用在因特网上。模糊值、模糊联系和语言量都通过 JAVA 来定义。

(4) Web 应用:通过 Web 日志的挖掘,来发现用户访问 Web 页面的模式。通过分析 Web 日志记录中的规律,可以识别电子商务的潜在客户,增强对最终用户的 Internet 信息服务的质量和交付,并改进 Web 服务器系统的性能。还可以进一步获得用户访问的附加信息(包括 Web 服务器缓冲区中用户浏览 Web 页面的序列等),以便于做更为详细的 Web 日志分析。如通过用户访问模式的学习改进其自身的 Web 站点,有助于建立针对个体用户的定制 Web 服务。为了挖掘出较完全的兴趣模式,吴瑞[9]提出一种新的结构类型——FLAAT,它可发现那些被忽略的用户浏览偏爱路径。同时引进模糊集来处理停留在网页上的时间,以形成语义术语使挖掘出的用户浏览偏爱路径更自然、更易理解。算法能准确地反映用户的浏览兴趣。

(5) 图像检索

随着近来由多种媒体数据构成的多媒体信息仓库数据的增加,基于内容的图像检索开始活跃在这个领域。和传统数据库中基于精确匹配的关键字来检索信息不同,基于内容的图像检索系统的信息是一个图像的可视特征。如颜色、纹理、形状等。由于检索中查询要求往往是根据人的主观性所决定,因此很大程度上带有模糊性。对于图像纹理,习惯于用"很粗"、"中等"、"弱"这样的一些模糊概念来描述;形状一般用"几何形的"、"立体形的"或"似长方形的"、"正方形的"等概念描述;颜色特征通常用"很艳"、"一般"、"暗淡"或"大红"、"紫红"、"红"这样的模糊概念来描述。所以基于内容是图像检索是基于图像的相似特征来检索的。

2.2 神经网络

数据挖掘的困难主要存在于三个方面:首先,巨量数据集的性质往往非常复杂,非线性、时序性与噪音普遍存在;其次,数据分析的目标具有多样性,而复杂目标无论在表述还是在处理上均与领域知识有关;第三,在复杂目标下,对巨量数据集的分析,目前还没有现成的且满足可计算条件的一般性理论与方法。研究者们主要是将符号型机器学习方法与数据库技术相结合,但由于真实世界的数据关系相当复杂,非线性程度相当高,而且普遍存在着噪音数据,因此这些方法在很多场合都不适用。

因为神经网络的黑箱问题,在数据挖掘的初期并不看好,然而,神经网络由于本身良好的鲁棒性、自组

织自适应性、并行处理、分布存储和高度容错等特性，以及它对未经训练的数据分类模式的能力，非常适合解决数据挖掘中存在的以上问题，因此近年来越来越受到人们的关注。

规则抽取方法是解决"黑箱问题"的有效手段。神经网络规则抽取的研究最早开始于 80 年代末。1988 年，Gallant[10]设计了一个可以用 if – then 规则解释推理结论的神经网络专家系统。根据设计思想的不同，目前的规则提取方法大致可以分成两大类，即基于结构分析的方法和基于性能分析的方法。

基于结构分析的神经网络规则抽取方法把规则抽取视为一个搜索过程，其基本思想是把已训练好的神经网络结构映射成对应的规则。由于搜索过程的计算复杂度和神经网络输入分量之间呈指数级关系，当输入分量很多时，会出现组合爆炸。因此，此类算法一般采用剪枝聚类等方法来减少网络中的连接以降低计算复杂度。RX 算法[11]首先用权衰减方法构造 BP 网络（该网络中连接权的大小反映了连接的重要程度），然后对网络进行修剪，在预测精度不变的情况下删除次要连接，在对网络进行充分简化的条件下，对隐藏层结点的激活值进行聚类，根据不同的隐藏层结点激活值用穷举搜索的办法来寻找从输入层到隐藏层和从隐藏层到输出层的规则。

与基于结构分析的方法不同，基于性能分析的神经网络规则抽取方法并不对神经网络结构进行分析和搜索，而是把神经网络作为一个整体来处理，这类方法更注重的是抽取出的规则在功能上对网络的重现能力，即产生一组可以替代原网络的规则。1994 年，Craven 和 Shavlik[12]为神经网络规则抽取任务下了一个定义：给定一个训练好的神经网络以及用于其训练的训练集，为网络产生一个简洁而精确的符号描述。在文献[12]的基础上，1996 年，Craven 和 Shavlik[13]提出了 TREPAN 算法。该算法首先用训练好的神经网络对示例集进行分类，然后将该集合作为训练集提供给决策树学习算法，从而构造出一棵与原网络功能接近的、使用 MOFN 表达式作为内部划分的决策树。TREPAN 的计算量较低。1997 年，Craven 和 Shavlik[14]将 TREPAN 用于一个噪音时序任务，即美元 – 马克汇率预测，取得了比现有方法更好的效果。

2.3 遗传算法

遗传算法是一种基于生物自然选择与遗传机理的随机搜索算法，是一种仿生全局优化方法。它是美国 Michigan 大学的 Holland 教授于 1975 年首先提出的。遗传算法中包含了 5 个基本要素：1）参数编码；2）初始群体的设定；3）适应度函数的设计；4）遗传操作设计；5）控制参数设定。遗传算法具有十分顽强的鲁棒性、自适应性，其在解决大空间、多峰值、非线性、全局优化等复杂度高的问题时具有独特的优势。因此，遗传算法在数据挖掘技术越来越显示出其重要的地位。遗传算法在数据挖掘中主要应用于数据回归和关联规则的发现。

（1）回归：除了发现可解释的模式之外，数据挖掘的另外一个重要的任务就是预测，即通过数据库中的一些变量发掘其超未来的趋势值。传统的线性回归需要先假设这些属性间没有相关性，而遗传算法则可以很好的处理有相关性的变量。Xu[15]曾设计了一个多输入单输出的系统，应用遗传算法从训练数据集中进行非线性多元回归。

（2）关联规则：遗传学习首先创建一个由随机产生的规则组成的初始群体。每个规则可以用一个二进制位串表示的 if – than 类型。通过全局搜索，形成由当前群体中最适合的规则组成新的群体。遗传算法可以单独用于数据仓库中关联规则的挖掘，还可以和其他的数据挖掘技术相结合，例如，用于进化神经网络结构以得到结构简单、性能优良的神经网络结构[16]；用于特征子集选择[17]；应用于决策树、分类器和模糊规则的获取等等。

2.4 粗集

粗集理论由波兰逻辑学家 Pawlak 教授在 20 世纪 80 年代提出，是一种处理含糊和不确定问题的新型数学工具。粗集理念基于给定训练数据内部的等价类的建立。给定现实世界数据，通常有些类不能被可用的属性区分。粗集可以用来近似定义这种类，将问题的数据集进行划分，然后对划分的每一部分确定其对某一概念的支持程度：即肯定支持此概念，肯定不支持此概念，并分别用下近似和上近似集合来表示为正域、负域。它能有效地分析不精确、不一致、不完整等各种不完备的信息，还可以对数据进行分析和推理，从中发现隐含的知识和潜在的规律。同时，粗集理论在处理大数据量，消除冗余信息等方面有着良好的效果，因此广

泛应用于数据挖掘的数据预处理、规则生成等方面。

(1) 数据约简:粗集理论可提供有效方法用于对信息系统中的数据进行约简在数据挖掘系统的预处理阶段,通过粗集理论删除数据中的冗余信息(属性、对象以及属性值等),可大大提高系统的运算速度。文献[18]使用粗集方法对信息系统进行属性及属性域的约简,然后使用神经网络对约简后的数据进行分类,从而在网络分类精度没有明显下降的前提下使网络的学习速度提高到约简前的 4.72 倍。

(2) 规则抽取:与其它方法(如神经网络)相比,使用粗集理论生成规则是相对简单和直接的,信息系统中的每一个对象既对应一条规则。粗集方法生成规则的一般步骤为:①得到条件属性的一个约简,删去冗余属性;②册去每条规则的冗余属性值;③对剩余规则进行合并目前已经产生了许多基于粗集理论的方法用于从信息系统中抽取规则[19]。

粗集理论存在对错误描述的确定性机制过于简单,而且在约简的过程中缺乏交互验证功能,因此,粗集理论与其它方法如神经网络、遗传算法、模糊数学、决策树等相结合可以发挥各自的优势,大大增强数据挖掘的效率。文献[20]提出了一种融合粗集理论和神经网络的数据挖掘新方法,应用于大型数据库的分类规则挖掘。其主要思想是首先由粗糙集理论对数据库进行初步约简,然后借助于神经网络在自学习过程中完成对数据库的进一步属性约简,并过滤数据中的噪声数据,最后由粗糙集理论对约简后的数据库进行规则抽取。粗集理论的使用提高了系统的运算速度,同时神经网络则使产生的规则集泛化能力提高。

2.5 混合方法

综合软计算的主要算法可产生在并行化、容错、自适应性和不定性管理方面更好的系统。混合系统可使许多应用中的自动化自适应系统成为现实。模糊系统的推理能力,当与神经网络和遗传算法的学习能力结合时,导致得到体现合理有效的认识系统(可学习和推理的系统)的新产品和新过程。Banerjee[20]利用粗糙集、神经网络和模糊逻辑相结合的方法设计了数据挖掘系统,其中用粗糙集方法在决策表中进行约简。而用模糊集方法挖掘出未经加工的知识,最后由神经网络根据依赖度进行取舍。

3 结束语

目前,数据挖掘中算法和可视化的研究越来越显得重要。因为从数据库中很容易就可以发现大量的模式,而这些模式中很多是很显而易见的、冗余的、无用的,或是对用户来说没有趣的。现在就需要能够过滤这些模式而提供给用户有用或有趣的模式的挖掘技术。软计算方法,包括模糊逻辑、神经网络、遗传算法、粗集和混合方法,近来用于解决这些问题。

软计算具有以低求解成本、快速的方法解决复杂问题。本文对数据挖掘中软计算方法及应用作了一个综合性阐述。对它们的特点进行了分析,并对它们在数据挖掘中的应用进行了分类。模糊集为这个过程中的处理不确定性提供了一个框架,神经网络和粗集广泛应用于分类和规则生成。遗传算法应用于各种优化和搜索过程中,如优化排序和模式选择。

参考文献

1 杨会志,数据挖掘技术的主要方法及其发展方向,河北科技大学学报[J],2000,21(3):77-80.

2 Zadeh L., Fuzzv logic, neural network and soft computing. Communications of the ACM, 1994, 37(3):77-84.

3 D. Nauck, "Using symbolic data in neuro-fuzzy classification," in Proc. NAFIPS 99, New York, June 1999, pp. 536-540.

4 汤效琴、戴汝源,数据挖掘中变量聚类方法的应用研究,计算机工程与应用[J],2004,40(24):171-173.

5 汤效琴、戴汝源,数据挖掘中聚类分析的技术方法,微计算机信息[J],2003(1):3-4.

6 Jawei han, 数据挖掘:概念与技术[M],北京:机械工业出版社,2001.

7 Q. Wei and G. Chen, "Mining generalized association rules with fuzzy taxonomic structures," in Proc. NAFIPS 99, New York, June 1999, pp. 477-481.

8 J. Kacprzyk and S. Zadrozny, "Data mining via linguistic summaries of data: An interactive approach," in Proc. IIZUKA 98, Fukuoka, Japan, Oct. 1998, pp. 668-671.

(下转第 114 页)

(上接第 122 页)

9 吴瑞,基于 FLAAT 模糊的 WEB 挖掘算法,武汉科技大学学报(自然科学版)[J],2005,28(3):270-272.

10 S. I. Gallant. Neural Nework Learning and Expert Systems. Cambridge, MA:MIT press, 1993.

11 Rudy Setiono, Liu H. Understanding neural networks via rule extraction. In: Proc of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995. pp. 480 - 485.

12 M. W. Craven, J, W, Shavlik . Using sampling and queries to extract rules from trained neural networks. In: Proc of the 7th Int1 Conf on Mathine Learning, New Brunswick, 1994. pp. 37 ~ 45.

13 M. W. Craven, J, W, Shavlik. Extracting tree - structured representations of trained networks. Cambridge, MA:MIT press, 1996.

14 M. W. Craven, J, W, Shavlik. Using neural networks in data mining. Future Generation Computer Systems. 1997. 13. pp. 211 - 229.

15 K. Xu, Z. Wang, and K. S. Leung, "Using a new type of nonlinear integral for multiregression: An application of evolutionary algorithms in data mining," Proc. IEEE Int. Conf. Syst., Man, Cybern., pp. 2326 - 2331, Oct. 1998.

16 郑志军、林霞光,一种基于神经网络的数据挖掘方法,西安建筑科技大学学报[J],2000,32.

17 刘勇国、李学明、张伟基,于遗传算法的特征子集选择,计算机工程[J],2003,29.

18 Jelonek J, Krawiec K. Rough set reduction of attributes and their domains for neural networks[J]. Computational Intelligence. 1995. 11(2):339 - 347.

19 Kryszkiewicz M. Rules in incomplete systems[J]. Information Sciences, 1999,113(4): 271 - 292.

20 Banerjee M. Pal K. Rough fuzzy MLP: knowledge encoding and classification[J]. IEEE Trans. Neural Networks, 2002. 9:1203 - 1216.