人工免疫系统在数据挖掘中的应用

徐春鸽,章登科,葛 红
(华南师范大学 计算机学院,广东 广州 510631)

摘 要:对自然免疫系统机制、人工免疫系统及数据挖掘进行简要介绍。对人工免疫系统在数据挖掘领域中的应用进行详细分析综述。主要阐述人工免疫系统在分类规则、聚类规则等中的应用现状,并对其方法进行详细分析并指出其优缺点。

关键词:人工免疫系统;数据挖掘;聚类规则;分类规则
中图分类号:TP18;TP301 **文献标识码**:A **文章编号**:1673-629X(2007)04-0034-04

The Application of Artificial Immune System to Data Mining

XU Chun-ge, ZHANG Deng-ke, GE Hong
(Computer College of South China Normal University, Guangzhou 510631, China)

Abstract: The natural immune system (NIS), the artificial immune system (AIS) and data mining are briefly introduced at first in the paper. Then it focuses on analyzing and detail survey of the application of artificial immune system to data mining. The application of artificial immune system on classification rule and clustering rule is mainly explained. At the same time, the methods are introduced in detail. The advantages and disadvantages of the application of artificial immune system to data mining are pointed out.

Key words: artificial immune system; data mining; clustering rule; classification rule

1 自然免疫系统

自然免疫系统是由体内循环的免疫细胞群构成的复杂分散没有中心器官的系统,从信息处理的观点看,免疫系统内部蕴含着信息的分布、识别、学习和记忆等复杂的信息处理机制:

(1)己与非己:免疫系统的反向选择原理表明,免疫系统能够识别体内分子与外来分子。

(2)学习与优化:免疫网络和克隆选择理论解释了免疫应答过程的机理,表明抗体的产生是免疫系统的学习过程。

(3)联想记忆:免疫系统消灭抗原后,产生记忆细胞,当与该细胞相似的抗原再次入侵肌体时,免疫系统能够产生更快速、更强烈的二次应答。

(4)自适应网络:免疫网络学说表明,免疫系统中的B细胞通过识别与被识别组成一个网络,其与神经网络一样,是一个能够学习与记忆的自适应网络。

(5)分布系统:免疫网络的淋巴细胞分布于全身,根据周围的环境自适应地确定自身的行为,是一个没有中心控制的并行分布自治系统[1]。

2 人工免疫系统

目前许多研究人员已经利用免疫系统某一个或某些性质、功能、机制发展出各种人工免疫系统技术,这些免疫技术可以被统称为人工免疫系统。

人工免疫系统包括算法、数学模型、混合智能系统等,大体可分为两类:

一类是从生物学角度模拟研究人工免疫系统。例如:Jamey提出的人工免疫网络理论,阐述了免疫系统具有大量独特型及抗体独特型的抗体,由此形成了细胞间相互制约关系,而基于免疫系统中B细胞和T细胞之间的相互反应可被借鉴用于建立人工免疫网络模型。aiNet网络,该模型通过克隆选择控制网络细胞数量和位置,最小生成树来定义最终网络结构,应用在数据聚类方面具有很好的自动联想记忆能力、概括能力、噪声耐受等。基于免疫系统检测抗原原理用于计算机安全的ARTIS[2,3]系统、有限资源人工免疫系统(RLAIS)[4~6]等等。

另一类是从研究免疫系统某些机制出发,从而加深对免疫系统的理解,对建立某些工程问题的人工免疫系统很有启发。例如:抗体基因库模拟进化,系统多

收稿日期:2006-07-02

作者简介:徐春鸽(1979-),女,河北南宫人,硕士研究生,研究方向为人工智能;葛 红,博士,副教授,研究方向为人工智能、自动控制。

样性产生[7]等。人工免疫系统已经应用于不同的工程领域,目前主要应用成果集中在异常和故障诊断、计算机网络安全和病毒检测、机器学习图像处理、自动控制等领域。

3　人工免疫系统在数据挖掘中的应用

3.1　数据挖掘

数据挖掘是一种特定应用的分析过程,可以从大量的数据中提取隐含的、未知的、潜在有用的,并且最终能被理解的模式。它本身是一个发展历史较为悠久的领域,并且也形成了一些比较有效的数据分析技术和方法。但是随着社会和科技的发展,产生大量的、复杂的、多样的数据,尤其是 Internet 的流行,出现了海量的、异构的半结构化的数据;随着移动通信和无线技术的发展出现了大量的移动的、动态的数据等等,使用原来的方法效率不高、收敛速度慢,甚至由于传统方法的缺陷根本不能解决问题等,可见传统数据挖掘方法已不再能满足人们的需求。因此人们不断地探索将技术新成果应用于数据分析领域,例如人工智能、遗传算法、神经网络等技术在数据分析技术中的应用以及数据挖掘概念的提出和研究,使得数据分析领域的研究取得了重大进展。

3.2　人工免疫系统与数据挖掘

数据挖掘从狭义上讲就是信息处理问题,很多都属于模式识别所研究的范畴。模式的分类和识别可以由分类器来实现。为了设计分类器必须首先对分类器进行训练,即分类器首先要进行学习,从而,使分类器具有自动识别能力。分类器的训练/学习方法又可分为两种,一种叫做预分类的训练试验,即监督训练和未分类的训练试验,即无监督训练。数据挖掘中的聚类与分类从训练学习的角度来看属于监督训练和无监督训练。而人工免疫系统在训练学习等方面具有其它系统无可比拟的优点,于是人们便开展了人工免疫系统应用于数据挖掘的研究。

人工免疫系统作为一种新兴智能系统,在数据挖掘中的应用刚刚起步,可以作为一种新的数据挖掘方法,来对数据库、数据仓库、文本、Web 页等进行数据挖掘,发现有用的知识。下面从聚类、分类的角度来阐述人工免疫系统的数据挖掘应用。

3.3　聚　类

人工免疫系统在数据挖掘中的应用目前更多是用在无监督学习即聚类方面,下面介绍人工免疫系统应用于聚类分析时的一般步骤及 RLAIS 系统和 aiNet 聚类算法两种具有代表性人工免疫聚类系统的具体应用过程。

人工免疫系统应用于聚类分析的一般步骤:首先初始化,建立最初的人工免疫网络,随机产生抗原并产生各种参数,例如:亲和力阈值等。然后利用抗原训练网络并计算亲和力,根据亲和力进行克隆选择和变异,并且调整网络,这样不断地训练学习直到满足结束条件,达到要求为止。

3.3.1　RLAIS 系统

Timmis 通过改进 Jisys 系统,提出了一种新的基于免疫网络的数据分析方法——RLAIS 系统。RLAIS 系统是基于免疫系统内的资源是有限的,这种限制导致 B 细胞之间的竞争,受刺激的最强的细胞才能存活的思想建立的。首次引入了人工识别球(ARB)的概念,来表示许多同样的 B 细胞。RLAIS 是由一组 ARB 和它们之间的联系组成,在学习训练开始,由该算法中的克隆和变异产生多样性。并将以前未见过的数据不断提交给网络,通过学习过程可以融合进网络,这样随着网络的不断学习,新的、好的、刺激度高的 ARB 会被保留,刺激度低的会被从网络中删除,直到满足结束条件为止。算法流程图如图 1 所示。

RLAIS 是针对聚类分析问题而设计的,应用效果也不错,是一种比较好的聚类分析新模式。但是,RLAIS 系统同样存在几个比较突出的问题:

(1)聚类分析结果意义不明确。对一组数据的聚类分析的结果可以是简单的对原始数据的聚类,可以是对原始数据组中所包含的模式类的表达。RLAIS 系统的结果是网络结构稳定,而此时的网络所代表的含义不清楚。

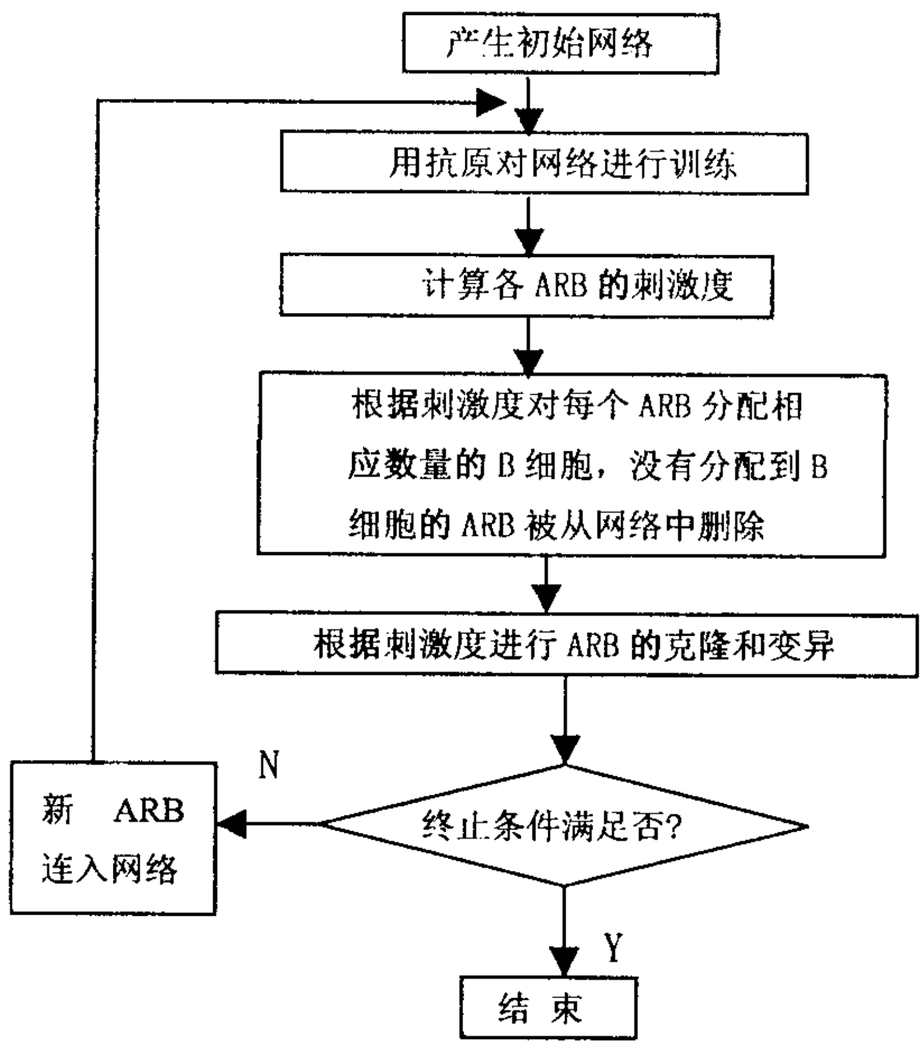

图 1　RLAIS 算法流程图

(2)算法中关于抗体的评价公式太简单。抗体与抗原的匹配度、抗体与抗体的刺激值和抑制值都用欧

氏距离度量,所有这些值都采用直接加、减的方式组合,这样的抗体评价值可能影响到整个系统的性能。

(3)亲和力成熟过程中,单纯的克隆选择和随机变异操作。完全的随机变异,虽可以在群体中引入新的抗体,但同时也会影响群体的收敛性和稳定性。

(4)记忆库的使用。记忆库未被用于学习过程中,这同样不利于系统的收敛速度和效果。

(5)未考虑用聚类准则函数来评判聚类结果的好坏。

3.3.2 aiNet 的聚类算法

aiNet 是由 De Castro 提出的,它是一个边界加权图,无需全部连接,又被称为细胞的节点集合组成,每一个连接边界具有一组分配权或连接强度[1]。该学习算法的目的是建立一个记忆集合,识别和表示数据结构组织。细胞越特异,网络简化程度越低(低压缩率),同时细胞越"通用",网络越简化(改善压缩)。抑制阈值 σ_s 控制细胞特异水平、聚类准确性和网络可塑性。作者对该算法进行了测试,虽然该算法是一般性的,但是结果网络与问题有关,主要缺点是定义参数多,计算成本高,网络对抑制阈值 σ_s 敏感等等。因此后来又有很多人对其进行了改进。aiNet 的优点是:对二、三维数据通过形成的免疫记忆数据以可视化效果反映原始数据之间的聚类结构,通过抑制阈值参数调节控制生成的记忆细胞数目;产生的记忆细胞质量较高。但其缺点是最终形成的聚类结果并不是原始数据的聚类,而是通过其算法形成记忆细胞,再利用图论和传统分级聚类方法间接反映原始数据中的聚类信息,不能给出原始数据的准确类别,而且只用于简单的二类问题,数据规模小,对四维以上数据聚类无法实现网络可视化等。有关研究人员对 RLAIS 和 aiNet 进行了不同形式的改进,如 Fuzzy RLAIS[8],SSAIS 等等[9]。

3.4 分 类

分类技术从机器学习的角度看属于监督学习技术,人工免疫系统在监督学习领域已有许多应用。基于免疫原理开发的监督学习算法包括 Immuno - 81[10],该系统模拟 B 细胞和 T 细胞性质,以及它们产生初次和二次免疫应答的相互作用机制等几个方面。Potter 描述使用了协同进化遗传算法 AIS 模型进化抗体,用于概念学习[11,12]。后来出现的 AIRS[13](资源有限人工免疫分类器)是一种主要监督学习算法,得到了广泛认可。AIRS 是建立在 Timmis 的资源有限人工免疫算法(RLAIS)基础上的。

3.4.1 AIRS

算法过程分为四步:标准化和初始化;产生 ARB 后代;资源竞争;保留更适合的记忆细胞。

1)初始化。

特征向量标准化,计算亲和力阈值,从训练数据集中随机选择种子记忆细胞群和 ARB 群。

2)产生 ARB 后代。

找出和抗原同类的记忆细胞和抗原最为匹配的记忆细胞,并产生新的克隆和变异的 ARBs, 加到 ARB 集。

3)竞争。

AB = MCmatch 所有新的克隆个体 + 以前所有对抗体反应中所剩的 ARBs,以上的这些组成分将会以当前抗原的刺激程度为基准进行资源竞争。

4)把具有更好性能的候选记忆细胞加入到记忆细胞池中。

如果 MCcand > MCmatch,它将被加入到记忆细胞池中,而且如果 MCcand 和 MCmatch 之间的距离小于亲和力阈值乘以亲和力阈值参数,MCcand 将取代 MCmatch,MCmatch 被删除。

该系统的特点是:

(1)提出了人工识别球(Artificial Recognition Ball, ARB)的概念,即一个 ARB 表示某种数量的 B 细胞或资源,系统中总的资源数量是有限的。

(2)抗体和抗原都是特征空间中的特征向量。抗体和抗原的亲和力是用 ARB 抗体和抗原之间的欧氏距离来确定的。

(3)系统中的资源总数是有限的。

(4)被激活的 ARB 进行克隆和变异产生后代。

(5)变异的后代也有机会竞争资源。

(6)和 B 细胞相比,记忆细胞存活的时间要长。

(7)由于 一些 ARBs 得到了新的资源会导致另外的 ARBs 从系统中删除。

由此可知该系统中应用了细胞亲和力、克隆和变异、免疫记忆、免疫网络等多种免疫原和概念,并在 UCI 数据集合上做了分类测试,收到了良好的效果。

3.4.2 AINMC 算法

2004 年莫宏伟等人利用 aiNet 产生免疫记忆的机制提出一种新型分类器系统 AINMC[11](人工免疫网络记忆分类器),算法流程图 2 所示。

该算法与 AIRS 的主要区别是:

(1)AINMC 的免疫记忆细胞是经过免疫细胞之间刺激和抑制过程之后产生的,也就是模拟免疫网络中 B 细胞相互作用机制;而 AIRS 则是通过比较提呈抗原和产生相应记忆细胞之间及与该抗原和记忆细胞池中的细胞之间的亲和力后产生的最终的记忆细胞。

(2)二者限制克隆的机制和过程有很大不同。AIRS 利用资源有限机制克隆规模;而 AINMC 则采用

网络阈值机制限制克隆变异细胞的规模。

(3)记忆细胞没有冗余,减少计算负担。

4 小　结

由上述的综述分析可知尽管在数据挖掘领域人工免疫系统取得了一定的成就,但是还处于起步阶段,还有很多缺点和需要改进的地方,现提出人工免疫数据挖掘技术研究方向如下:

(1)改进和完善上述方法。由上述分析可知许多方法还存在缺陷,需要对其加以改进,尤其是人工免疫系统在关联规则中的应用还很少,因此更待加强。

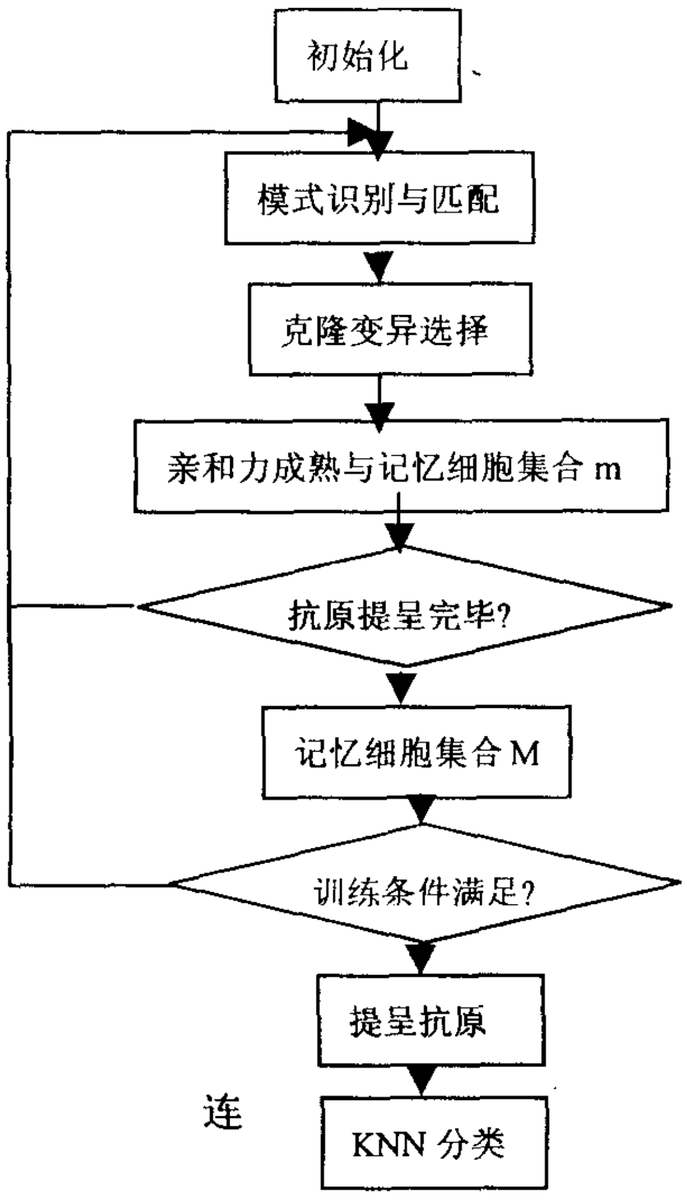

图 2　AINMC 算法流程图

(2)与现有技术进行比较研究。针对上述现有数据挖掘技术存在的问题,将基于免疫原理的数据挖掘技术与数据挖掘技术进行比较研究。通过比较找到现有技术和人工免疫数据挖掘技术优缺点,对人工免疫系统技术加以改进,以互相融合应用。

(3)人工免疫数据挖掘技术的免疫学原理研究。在已开发的用于数据分析的人工免疫系统中,B 型淋巴基本目的是用于机器学习,由于 B 和 T 细胞相互作用的复杂性,一般忽略 T 细胞的辅助模式识别作用。B 细胞群体和 T 细胞群体相互作用实现模式识别任务的原理和机制可以作为一个研究重点[11]。

(4)人工免疫数据挖掘的数学基础。目前人工免疫系统对免疫系统机制依赖过多,还没有形成严格的数学理论,多数算法缺乏数学基础等。

(5)进一步拓宽人工免疫数据挖掘在工程领域中的应用。例如数据挖掘中新的热点 Web 知识发现,目前人工免疫系统在 Web 知识发现中的应用还很少,所以无论从技术还是商业应用角度来说在这个方向是很有潜力和前景的。

参考文献:

[1] 莫宏伟.人工免疫系统原理及应用[M].哈尔滨:哈尔滨工业大学出版社,2002:24-47.

[2] Hofmeyr S A, Forrest S. Immunity by Design: An Artificial Immune System[C]//Proc of GECCO'99. USA: [s. n.], 1999:289-296.

[3] Hofmeyr S A, Foreest S. Architecture for an Artificial Immune System[C]//submitted to Evolutionary Computation. [s. l.]: [s. n.], 2000.

[4] Timmis J, Neal M, Hunt J. An artificial immune system for data analysis[J]. Biosystems, 2000, 55(1/3):143-150.

[5] Timmis J, Neal M. A resource limited artificial immune system for data analysis[J]. Knowlege-Based System, 2001(14):121-130.

[6] 葛　红,毛宗源.免疫算法及核聚类人工免疫网络应用研究[D].广州:华南理工大学自动化与工程学院,2003.

[7] Bernaschi M, Castiglione F, Sucdi S. A High performance simulator of the Immune Response[J]. Future Generation Computer Systems, 1999, 15:333-342.

[8] Hightower R H, Forrrest S, Perelson A S. The Baldwin Effect in the Immune System: Learning by Somatic Hypermutation [C]//Belew R K, Mitchell M. In Adaptive Individuals in Evolving Populations, Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley, 1996:159-167.

[9] 莫宏伟,吕淑萍.基于人工免疫网络的新型分类器研究[J].计算机工程与应用,2004(36):28-32.

[10] Carter J H. The Immune System as a Model for Pattern Recognition and Classification[J]. The American Medical Informatics Association, 2000, 7(1):28-41.

[11] 莫宏伟,吕淑萍.基于人工免疫系统的数据挖掘技术原理与应用[J].计算机工程与应用研究,2004(14):28-32.

[12] Potter M A, De Jong K A. The convolution of antibodies for concept learning[C]//In: Eiben A E, Back T. Proc of the Fifth Int Conf on Problem Solving from Nature(PPSN-98). Berlin: Springer-Verlag, 1998.

[13] Watkins A, Boggess L. A New Classfier Based Resource Limitied Artificial Immune Systems[J]. Proceedings of IEEE, 2002, 2(4):82-90.