混合智能系统 R-CSNN 及其应用*)

夏 洁[1] 蔡自兴[1] 王 刚[2]

(中南大学信息科学与工程学院 长沙 410083)[1] (复旦大学管理学院 上海 200433)[2]

摘 要 常见分类算法面对海量数据时,在时间效率、鲁棒性和精确性上都显示出了不足。为此,本文将混合智能系统引入到神经网络分类算法的研究中,针对传统神经网络算法在训练速度和鲁棒性上的不足,提出了一个基于粗糙集、免疫算法和神经网络的混合智能系统。该系统以粗糙集为前端处理器,在保证一定信息量的基础上对输入到神经网络的数据进行约简,接着用改进免疫算法作为学习算子的神经网络进行训练。最后,通过 UCI 下的数据库进行仿真实验,验证了该系统的有效性。

关键词 粗糙集,人工神经网络,免疫算法

Hybrid Intelligent Systems R-CSNN and its Applications

XIA Jie[1] CAI Zi-Xing[1] WANG Gang[2]

(Information Engineering School, Central South University, Changsha 410083)[1] (Management School, Fudan University, Shanghai 200433)[2]

Abstract Facing the huge amounts of data the familiar classification algorithms show the shortages on the time efficiency, robustness and accuracy. So this article puts the hybrid intelligent systems into the research of classification algorithm. We propose hybrid intelligent systems based on rough set, immune algorithm and neural networks aiming at traditional neural networks algorithm's shortage on training speed and robustness. The system takes rough set as a front processor, and reduces the data which are input into neural networks on the basis of preserving definite information amount, then trains the neural networks using an improved immune algorithm as operators. Finally, experiments are carried out based on the data of database. It is observed that the system is valid.

Keywords Rough set, Artificial neural network, Immune algorithm

1 引言

自从 Minsky 等学者在 1956 年提出人工智能的概念以来,人工智能已经走过了五十年的研究历程,并取得了一些划时代的成果。这些研究成果为人工智能的发展奠定了基础,丰富了人工智能的研究内容。人工智能这一智能模拟学科仍在发展,在其发展过程中还面临一些基本问题有待解决,主要有:(1)包含许多知识的人工智能系统的实时性问题;(2)对环境中的不完全的、模糊的、甚至部分错误的信息处理问题;(3)知识自动获取问题。针对 AI 研究中遇到的这些问题,目前智能系统研究的焦点很大一部分集中在将多种智能技术综合的研究上,也即形成了 AI 领域的一个新的研究方向——混合智能系统(Hybrid Intelligent Systems)。

混合智能系统是利用各种知识表达模型的不同特性,综合多种智能系统来对同一个事物进行多方面(多维)描述,以提高人工智能系统的性能(智能度、准确度),利用多维知识表达处理模型的思想来模拟人类智能行为,其目的是使建立的系统在知识表示、推理等方面更有效。早在 1991 年,AI 领域的著名专家 Minsky 就认识到研究不同智能技术组成的人工智能系统的必要性[1]。20 世纪 90 年代初,钱学森教授也提出了综合集成研究更是将机器体系、专家体系和知识体系有机结合起来而构成的一个高度智能化的人机结合系统。

将混合智能系统引入分类算法的研究,可望使分类算法在速度、精确度、鲁棒性等方面得到改善。正是按照这个思路,本文设计了一个用于分类的粗糙集、神经网络、免疫算法的混合智能系统。从基于生物体系发展的智能系统角度看,免疫计算智能和其它智能方法,如 ANN、EC、FS、DNA 计算等有紧密联系,可以相互融合利用发展新型计算智能系统和理论。本文旨在分析免疫优化算法特点和神经网络的基础上,提出改进算法。

2 模型设计

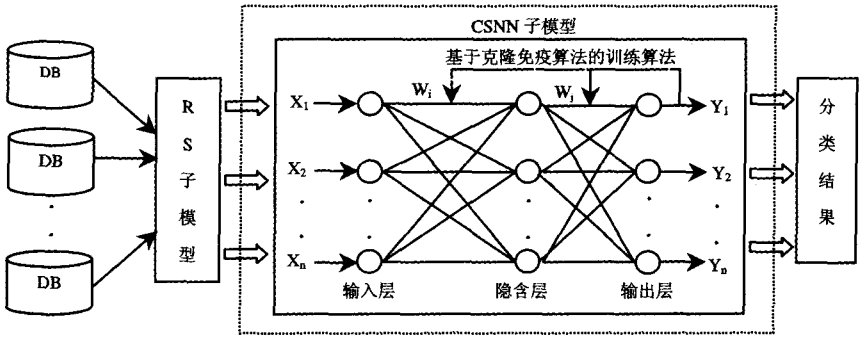
2.1 混合智能系统 R-CSNN 的总体框架

人工神经网络与免疫算法都是受生物学启发而来的理论和技术,两者在生物学原理和人工原理上有许多相同点和不同点,研究表明,免疫原理可以应用到神经网络来提高神经网络的性能;反之,神经网络理论也可以应用到免疫算法中,从而产生了一个相互之间可以受益的研究领域。粗糙集(Rough Set, RS)理论是 20 世纪 80 年代由 Pawlak 提出的一种处理模糊性与不确定性的数学工具[2]。在处理大量数据,消除冗余信息方面,粗糙集理论有着良好的结果。但是,由于粗糙集理论对错误信息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的功能,因此其结果往往不稳定、精度不高。粗糙集理论的知识约简方法可以利用并行算法实现,神经网络的最大特点之一就是无须实现函数表达而完成并行处理,且有容错和抗干扰的能力[3]。

由此可见神经网络、免疫算法和粗糙集即各有特点,又具有很多共同之处,探索三者的有机结合,可望为智能信息处理开拓一个光辉的前景。

综合粗糙集、神经网络和免疫算法的特点,首先通过粗糙集将输入数据进行删减,然后通过免疫算法改进了的神经网络进行训练,从而完成分类的整个训练过程。具体框架见图 1 所示。

*)基金项目:国家自然基金项目(编号:60404021);**蔡自兴** 教授,博士生导师,纽约科学院院士,联合国专家,CCF 高级会员,主要研究领域:人工智能,智能控制,智能机器人和自动控制等。**夏洁** 硕士研究生,研究方向:人工智能,模式识别与智能系统,智能 CAI。



DB-数据库；RS-粗糙集；CSNN-改进的免疫算法优化神经网络

图 1　混合智能系统 R-CSNN 总体框架

2.2　基于粗糙集的数据约简

粗糙集理论是一种新型的处理模糊和不确定知识的数学工具，目前已成为机器学习、数据挖掘等方面的新方法[4]。其中，由于属性约简是 NP(Nondeterministic Polynomial)问题，因此成为粗糙集理论研究中的重点问题之一。粗糙集理论中的知识表达方式一般采用信息表或称为信息系统的形式。信息系统可用四元有序组 $K=(U,A,V,\rho)$ 表示，其中 U 是对象的全体，即论域；A 是属性全体；$V=\bigcup_{a\in A}V_a$，V_a 属性的值域；$\rho:U\times A\to V$ 是一个信息函数，$\rho_x:A\to V$，$x\in U$，反映了对象 x 在 K 中的完全信息 $\rho_x(a)=\rho(x,a)$[5,6]。

对于这样的信息系统，每个属性子集都定义了论域上的一个等价关系，即 $\forall a\in A$ 决定等价关系 R_a，对属性集 $B\subseteq A$ 决定等价关系 $R_B=\bigcap_{a\in B}R_a$：$\forall x\in U,[x]_{R_B}=\bigcap_{a\in B}[x]_{R_a}$。

属性集 A 还可分为条件属性 C 和决策(结论)属性 D，这时的信息系统也称为决策表，常记为 $(U,C\cup D,V,\rho)$。

定义　设 $K=(U,A,V,\rho)$ 是一个信息系统，由 $B\subseteq A$ 所导出的等价关系为 R_B

①设 $a\in A$，若 $R_A=R_{A\setminus\{a\}}$，则称属性 a 是多余的；

②若在系统中没有多余属性，则称 A 是独立的；

③子集 $B\subseteq A$ 称为是 A 的约简，若 $R_B=R_A$，且 B 中没有多余属性，常记作 $red(A)=A$ 的全体约简；

④A 的所有约简的交集称为 A 的核，记为 $core(A)$。

对于给定的信息表，可以利用属性重要性来分析表中不同属性(因素)对决策属性的依赖程度。特别是应用属性约简的手段，可以剔除数据中存在的冗余成分和不相容性[7]，提取关键信息，生成决策规则，从而为预测和决策提供支持。在粗糙集理论中有一种生成对称矩阵的约简算法，其具体步骤如下[8]：

Step1：由矩阵的定义建立起矩阵，根据决策表的一致性定义判断决策表的一致性并进行分类。若表不一致，可将决策表分为两个子表：一致决策表和不一致决策表，然后将一致决策表按属性归类。

本方法的矩阵的每一项都分为两个小项：$c_{ij}1$ 和 $c_{ij}2$。这样可以极其方便地分析数据的变化。

其中：当 $(\exists m,n\wedge m\neq n\wedge u_i(m)\neq u_j(m)\wedge u_i(n)\neq u_j(n))$ 时，$c_{ij}1=\infty$；

当 $\forall m(m\in c\wedge u_i(m)=u_j(m))$ 时，$c_{ij}1=0$；

当且仅当 $m\in c\wedge u_i(m)\neq u_j(m)$ 时，$c_{ij}1=m$；

当 $\forall m\in D,u_i(m)=u_j(m)$ 时，$c_{ij}2=0$；

当 $\exists m\in D,u_i(m)\neq u_j(m)$ 时，$c_{ij}2=1$；

在这里：i,j 是布尔表中的对象，m,n 是指表中的属性，$u_j(m)$ 则是指表中的属性值。

Step2：计算决策表的核集 $core(c)$。

若 $\exists m\in C$，且 $D_m=\varnothing$，那么认为 m 为非核属性。

若 $\exists m\in C$，且 $D_m\neq\varnothing$，同时所有在 D_m 中的 $c_{ij}1=0$，则 m 为可约属性。

若 $\exists m\in C$，且 $D_m\neq\varnothing$，同时在 D_m 中存在 $c_{ij}2=1$，则 m 为核属性。

其中，D_m 的值定义如下：若 $c_{ij}1=m(m\in C)$，则加入属性表 D_m 中，即 $D=\{\}+(u_i,u_j,m)$。

这里我们使用上述算法，在保证整体分类精度的前提下对原始数据进行约减，以提高神经网络的学习效率。

2.3　改进的免疫算法优化神经网络(CSNN)模型

人工神经网络、GA 等优化方法的求解思想本质上都是“好上加好”，即不断向“优解爬山”来得到最终解，其局部和全局搜索能力在很多情况下都不能令人满意，在 GA 中增加免疫算子的作用也是有限的。近年来，人们意识到生物免疫机制对开发新的计算智能的启示意义，基于人工免疫的计算模型及应用已开始成为当前的一个研究前沿。根据生物免疫学的克隆选择原理，以克隆选择算子为核心，结合免疫反应中抗体群体多样性机制提出的免疫克隆算法如图 2 所示[9]。

算法：CSA 算法
输入：训练样本 samples
输出：训练样本集
方法：
①随机产生初始抗体群体，规模为 n，将抗体群体构成初始记忆细胞集合；
②根据 $F(P)=r_P(U,D)+\left[\frac{N-m}{N}\right]$ 的适应度函数计算每个抗体的适应度；
③对每个抗体进行克隆，产生一个克隆群体；
④对克隆群体进行高频变异，产生一个成熟的抗体群体，其中变异率与抗体的适应度成比例，同时还将保留原抗体，目的是为了防止变异后的抗体性能退化；
⑤计算成熟抗体群体中抗体的适应度；
⑥分别在 n 个抗体及其克隆群体中选择各自适应度最高的的抗体，并计算该克隆的平均适应度，将选择的 n 个最高适应度的抗体析的抗体群体；
⑦若平均适应度显著变化，则返回(2)继续优化，否则，将亲和力相近的抗体进行压缩，并剔除结构相同的抗体；
⑧在抗体群体中引入新的随机初始化抗体，以维持抗体群体的多样性；
⑨将新生成的抗体群体作为新的记忆细胞，返回(2)继续进行进化，直到满足停止条件。

图 2　克隆免疫算法

在上述算法中，(2)到(6)描述了对全体抗体进行克隆，模拟了免疫应答的分布性，通过与适应度成比例的变异实现了抗体局部最优的进化。在(7)和(8)中，当抗体群体趋于稳定状态时(由平均适应度衡量)，通过对相似的抗体进行剔除，来保持抗体的浓度。同时，对抗体群体增加一定的随机抗体，维持了抗体群体的多样性，从而可获得稳定的抗体局部优化解。

克隆选择是免疫优化的重要方式，在人工免疫系统中被广

为应用。克隆选择算法(Clone Selection Algorithm, CSA)模拟免疫细胞克隆选择原理。被选择的细胞受制于亲和力成熟过程,该过程改善对抗原的亲和力[9]。克隆选择免疫算法是一种收敛很强的并行算法,可以应付入侵抗原的变种,从而进一步提高免疫反应的识别多样性,有助于防止陷入局部最优解和优化早熟收敛,但并行算法都不可避免地存在着局部优化较弱的缺点。相反,神经网络具有局部寻优强而全局性较差的特点,根据 GA-BP 算法的思路,提出了 CS-LM 算法,即先用全局性强的 CS 算法进行搜索,然后在此基础上用 LM 算法进行局部寻优。这种算法结合了并行优化算法 CS 和局部优化算法 LM,具体流程如图 3 所示,LM 算法请详见文[10]。

3 仿真实验

为了验证混合智能系统 R-CSNN 的有效性,我们使用 UCI 机器学习数据库中的 6 个数据集作为实验数据集,各数据集的基本信息如表 1 所示。实验环境为 600Mhz CeleronII CPU,128MB 内存,操作系统为 Microsoft Windows2000,编程软件为 Matlab6.5。

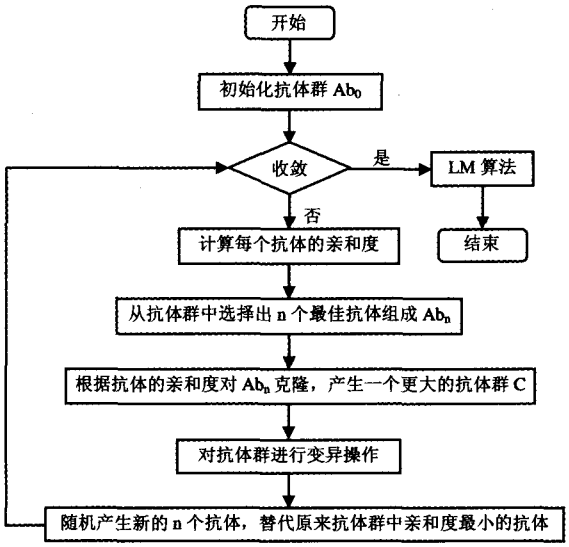


图 3 改进的免疫算法优化神经网络(CSNN)模型

表 1 数据集概况

数据集	属性数	分类数	样本数
BREAST	9	2	683
CAR	6	4	1728
CMC	9	3	1473
IRIS	4	3	150
YEAST	8	10	1484
ZOO	8	7	101

对各实验数据集采用混合智能系统 R-CSNN 训练时,对 CSNN 子模型中神经网络结构的确定,这里我们采用 3 层结构的网络,隐含层节点个数根据经验公式 $n1=\sqrt{m+n}+\beta$ 确定,其中 m 为输出层节点数,即为各数据集的属性数,n 为输入层节点数,即为各数据集的分类数,β 为 1－10 之间的常数;在克隆免疫算法中,种群大小 Ng=100,选择亲和度最高的 10 个体,复制大小 Nc/Ng=10,每一代引进的新个体数量 $d=10$,参数 $\beta=100$,迭代的代数=200;CSNN 子模型中神经网络的误差值我们统一取 0.1(接下来的对比实验中,为了有可比性,各神经网络的误差值也统一取 0.1)。关于训练集和测试集的获得本文采用保持法,即随机将原始数据集 2/3 的数据用作训练集,剩下的 1/3 用作测试集。限于篇幅,具体的训练过程这里略去。

我们将提出的分类算法与经典的 BP 算法、LM 算法进行了比较,其中各算法的参数取值同混合智能系统 R-CSNN 一致,30 次重复实验的平均结果如表 2 所示。其中 BP,LM 表示分别采用 BP,LM 算法的神经网络,R-CSNN 表示混合智能系统 R-CSNN。对每一种算法,表中第一列数据为分类的精确百分度,第二列数据中"㊣"表示实验过程中网络都收敛,"?"表示实验过程中网络存在发散的情况,"-"表示在此精度下,网络无法收敛。

表 2 混合智能系统 R-CSNN 模型与经典算法比较

数据集	BP		LM		R-CSNN	
BREAST	70.66%	?	73.33%	?	73.33%	㊣
CAR	-	-	67.28%	?	70.66%	㊣
CMC	-	-	63.57%	?	66.49%	㊣
IRIS	76.67%	㊣	80.00%	㊣	80.00%	㊣
YEAST	-	-	66.24%	?	69.14%	?
ZOO	81.26%	㊣	83.66%	㊣	83.66%	㊣

通过对比实验我们不难看出,随着样本数量的增加,原有的 BP、LM 算法都会出现不能收敛的情况,但混合智能系统 R-CSNN 则在保证了一定的分类精度的条件下基本上都能收敛,这就使得算法在鲁棒性和精确性上得到了提高。

结束语 从大量的观察和实验数据获取知识、表达知识、推理决策规则是智能信息处理的重要任务。本文设计的用于分类的粗糙集-神经网络-免疫算法混合智能系统(R-CSNN)的主要优点有:(1)利用粗糙集理论化简样本及条件属性,使得神经网络的输入样本大大减少,简化了神经网络结构,增强了系统的鲁棒性;(2)克隆选择算法是从许多初始点开始并行搜索,而不是从一个点开始,不仅搜索效率高,而且可以有效地防止搜索过程收敛于局部最优解;(3)基于 LM 算法的神经网络是在克隆选择算法搜索的基础上进行的搜索,不仅"少走了很多路",提高搜索效率,并且可以避免陷入局部最小。

然而我们同时也要看到,粗糙集-神经网络-免疫算法混合智能系统也还有很多问题要继续研究,比如各子模型参数的选取,子模型对整体性能的影响等等。但是我们相信,随着对混合智能系统研究的不断深入,上述问题一定能够解决,人类获取知识、表达知识和进行智能推理的愿望一定能够实现。

参 考 文 献

1 Minsky M. Logic versus analogical or symbolic versus connectionist or neat versus scruffy [J]. AI Magazine, 1991, 15(2):81～89
2 刘清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001
3 Pawlak Z. Rough Sets [J]. Communications of ACM, 1995, 38(11):89～95
4 梁霖,徐光华. 基于克隆选择的粗糙集属性约简方法[J]. 西安交通大学学报,2005,39(11): 1231～1235
5 Pawlak Z. Decision Table Computer [J]. Bulletin of the Polish Academy of Sciences Technical Sciences, 1986, 34(10):591～595
6 Pawlak Z. On Superfluous Attributes in Knowledge Representation System [J]. Bulletin of the Polish Academy of Sciences Technical Sciences, 1984, 32 (3): 211～213
7 曾黄麟. 粗糙理论及其应用(修订版)[M]. 重庆:重庆大学出版社,1998
8 王钰,王任,等. 基于 Rough Set 理论的"数据浓缩"[J]. 计算机学报,1998,21(5): 393～400
9 De Castro L N, Von Zuben F J. The clonal selection algorithm with engineering applications [A]. In: Workshop Proceedings of GECC'00, Workshop on Artificial Immune Systems and Their Applications [C]. Las Vegas, USA, 2000. 36～37
10 王刚. 基于混合智能系统的数据挖掘分类算法研究[D]. 长沙:中南大学,2004