用广义选择性神经网络集成预测 MHC-II 分子结合肽

胡桂武
HU Gui-wu

广东商学院 数学与计算科学系,广州 510320
Department of Mathematics Computational Science,Guangdong Business College,Guangzhou 510320,China
E-mail:pophu998@sohu.com

HU Gui-wu.Using generalized selective neural network ensemble to predict MHC class-II binding peptides.Computer Engineering and Applications,2008,44(18):9-11.

Abstract: Predictions of the binding ability of antigen peptides to Major Histocompatibility Complex(MHC) class II molecules are important for immunology research and vaccine design.The variable length and other aspects of each binding peptide complicate this prediction.In this paper,generalized selective neural network ensemble is proposed for prediction of MHC class II-binding peptides,the ensemble is built on two-level ensemble architecture.The first-level ensemble is used to create primary Neural Network Ensemble(NNE),where differential evolution is used to build some NNEs.The second-level ensemble is that a subset of primary NNEs is selected to make up the final ensemble.Experiment results indicate that the generalized ensemble model has better generalization and performance compared to traditional selective neural network ensemble.

Key words: selective neural network ensemble;differential evolution;Major Histocompatibility Complex(MHC)

摘 要:MHC II 类分子结合肽的预测对于免疫研究和疫苗设计非常重要，然而其结合肽长度的可变性等原因使其预测变得极为困难,提出了一种基于广义选择性神经网络集成的 MHC II 分子结合肽预测算法,该算法是一种双层集成模型。第一层是用微分进化算法去生成初始神经网络集成池,第二层是从初始神经网络集成池中选择部分组成最终的神经网络集成。实验结果表明广义选择性神经网络集成比传统的选择性神经网络有更好的泛化性能。

关键词:选择性神经网络;微分进化算法;MHC
DOI:10.3778/j.issn.1002-8331.2008.18.003 **文章编号:**1002-8331(2008)18-0009-03 **文献标识码:**A **中图分类号:**TP301.06

MHC II 类分子结合肽预测是免疫信息学中的一个重要且复杂的问题,近年来,随着生命科学、计算机科学、数学等多学科交叉融合与发展,MHC II 类分子结合肽预测技术取得了迅速的进步,迄今为止,主要有基于基序[1]、定量矩阵[2]、结构[3]和机器学习[4]等 4 大类预测方法,其中基于基序的方法由于较难找到明确的基序,其预测精度比较低[1],基于定量矩阵的方法具有比较高的预测精度,但不能处理数据中的非线性问题而受到限制[2],基于结构的方法一个明显的不足是速度慢,另外由于结合肽数据三维结构的不确定性[3],实际上不可能找到一种 MHC 分子结合肽预测的通用方法,基于机器学习[4](特别是基于神经网络的方法)由于能处理复杂的非线性问题,且具有较强的泛化能力和自适应能力,预测精度比较高等特点,受到许多研究者的广泛关注。

尽管神经网络都能以任意精度逼近任意复杂度的函数,但具体操作起来却很困难,其泛化性能有很大的局限性,幸运的是,20 世纪 90 年代初,Hansen 和 Salamon[5]提出了神经网络集成范式,并证明了通过集成技术,能大大提高神经网络的泛化能力,目前该技术已在与分类和回归相关的语音识别、图像分类、股市预测等多个领域取得了相当成功的应用,然而问题是神经网络训练个数越多,集成的时间、空间等付出的代价也越大,为了解决集成技术的局限性,近年来周志华等提出了选择性神经网络集成范式[6],理论分析表明从已有的个体学习器中进行选择后再集成,就可以获得更好的性能。

作为以上工作的一个进一步深入,笔者提出了广义选择性神经网络集成新范式,该范式是一个双层集成结构,第一层是生成一系列的初始神经网络集成,第二层是从第一层中选择一部分再次集成,即广义选择性神经网络集成。本文的另外一个工作是提出了基于微分进化算法的选择性神经网络集成方法;最后用广义选择性神经网络集成预测了 MHC II 类分子结合肽。

1 广义选择性神经网络集成

1.1 微分进化算法

微分进化算法(Differential Evolution,DE)是在 D 维空间

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60574078);广东省自然科学基金(the Natural Science Foundation of Guangdong Province of China under Grant No.06301003)。
作者简介:胡桂武(1970-),男,博士,副教授,主要研究领域:智能计算、生物信息学等。
收稿日期:2008-02-25 **修回日期:**2008-03-21

中选取 N 个 D 维向量 $\boldsymbol{x}_{i,G}$（个体），i=1,…,N 作为一个种群，在整个进化过程中群体规模不变。差分进化也有类似遗传算法的变异、交叉和选择等操作。基本差分进化算法详细过程引用文献[7]：

步骤 1 随机生成初始种群，DE 依次对每一个个体进行如下操作。

步骤 2 通过变异操作产生变异向量 $\boldsymbol{v}_{i,G}$，生成的方法如下：

$$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{r1,G}+\eta\cdot(\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G}) \tag{1}$$

其中 $r1,r2,r3\in\{1,\cdots,N\}$ 为随机整数，表示个体在种群中的序号；$\eta\in[0,2]$ 为收放因子，其中控制了序号为 $r1$ 和 $r2$ 两个向量的差异向量的放大量。

步骤 3 经交叉操作生成的探测向量为 $\boldsymbol{x}'_{iG}=[x'_{1I},x'_{2I},\cdots,x'_{DI}]$，生成的方法如下：

$$x'_{ji}=\begin{cases}v_{ji} & \text{if } rb(j)\leqslant CR \text{ or } j=rd(i)\\ x_{ji} & \text{if } rb(j)>CR \text{ and } j\neq rd(i)\end{cases} \tag{2}$$

其中：v_{ji} 是 $\boldsymbol{v}_{i,G}$ 的第 j 个分量；x_{ji} 是 $\boldsymbol{x}_{i,G}$ 的第 j 个分量；$rd(j)$ 是在[0,1]之间的随机数，$rd(j)$ 是[1,n]之间的随机整数；CR 一般是[0,1]之间的随机数。

步骤 4 差分进化的选择操作对于最小化问题的定义如下：

$$\boldsymbol{x}_{i,G+1}=\begin{cases}\boldsymbol{x}'_{i,G} & \text{if } Fit(\boldsymbol{x}'_{i,G})\leqslant Fit(\boldsymbol{x}_{i,G})\\ \boldsymbol{x}_{i,G} & \text{otherwise}\end{cases} \tag{3}$$

其中 $Fit(X)$ 为向量的适应值函数。

步骤 5 通过以上 DE 的变异交叉和选择操作使种群进化到下一代，反复循环进化最后种群将达到最优。

从上述寻优过程可以看出，DE 进化的本质是利用了群体中向量（个体）的距离和方向信息。比较容易实现，目前 DE 发展很快，出现了不同的工作策略，一般用 *DE/x/y/z* 表示[8,9]其中，DE 是指差分进化算法，x 表示 DE 变异时是使用“rand”（随机个体）是”best”（最好个体），y 为差异向量的个数，z 代表交叉操作方案是“bin”（二项式）还是“exp”（指数）。其中二项式交叉在交叉时对 D 维空间的每一个变量生成随机数判断，而指数交叉在交叉时只在 D 维空间生成随机数。

1.2 基于微分进化算法的神经网络集成

假定已经分别训练出 n 个神经网络 $f_1,f_2,f_3,\cdots,f_n$，从集合 $\{f_1,f_2,f_3,\cdots,f_n\}$ 中选择适合组成神经网络集成的子集 S，可以用上述 DE 优化算法。令每一个个体代表 $\{f_1,f_2,f_3,\cdots,f_n\}$ 中的一个子集 S，个体长度（个体空间的维数）等于神经网络的数量 n，个体中每一维取值为离散的 0 或 1，若在某一维取值为 1，表示对应的网络个体参与集成，若为 0，则不参与。由此选择性神经网络个体构建神经网络集成的问题可以转化为在 n 维 0–1 空间选择个体的 DE 优化问题。

选择性神经网络集成的关键是如何选择部分个体，目前还没有确定的最佳方法，有代表性的方法是基于遗传算法[10]的选择方法，笔者在此用微分进化算法去生成一系列的神经网络集成的重要原因是：差分进化算法有不同的进化模式（见表 1），具体操作如下。

假设已独立训练出 N 个网络 $S=\{f_1,f_2,f_3,\cdots,f_N\}$，需要找到 S 的一个子集 S'，使得 S' 的集成具有最佳的泛化性能，因为需要估计集成的实际泛化性能，引入验证集 $V=(x_i,y_i)$，其中 y_i 为第 i 个样本的实际输出值，选择性集成要找到 N 个网络的某个子集，使其具有最好的泛化能力，直观的想法是对所有的子集的解空间遍历，这种方法理论上能找到最优解，但是 N 个网络的组合有 2^N 种，当 N 比较大时（N>30）其时间复杂性极高，微分进化算法具有全局寻优性质的优化算法，在变异、交叉选择等算子作用下，通过对群体的进化获得对于现实世界问题的较优解，具体操作如下：DE 算法中的个体 X 用二进制编码，$X=\{x_1,x_2,\cdots,x_N\}$，其中当 x_i=1 时，表示 S 中的第 i 个网络参加本次集成，当 x_i=0 时，表示 S 中的第 i 个网络不参加本次集成，则个体 X 对应 S 中的一个子集 S^x 其适应值定义为：

表 1 典型的差分进化模式

序号	DE 的工作模式	公式
1	*DE/best/1/exp*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{best,G}+F\cdot(\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G})$
2	*DE/rand/1/exp*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{r1,G}+F\cdot(\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G})$
3	*DE/randtobest/1/exp*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{i,G}+F\cdot(\boldsymbol{x}_{best,G}-\boldsymbol{x}_{i,G})+F\cdot(\boldsymbol{x}_{r1,G}-\boldsymbol{x}_{r2,G})$
4	*DE/best/2/exp*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{best,G}+F\cdot(\boldsymbol{x}_{r1,G}+\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G}-\boldsymbol{x}_{r4,G})$
5	*DE/rand/2/exp*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{r5,G}+F\cdot(\boldsymbol{x}_{r1,G}+\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G}-\boldsymbol{x}_{r4,G})$
6	*DE/best/1/bin*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{best,G}+F\cdot(\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G})$
7	*DE/rand/1/bin*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{r1,G}+F\cdot(\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G})$
8	*DE/randtobest/1/bin*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{i,G}+F\cdot(\boldsymbol{x}_{best,G}-\boldsymbol{x}_{i,G})+F\cdot(\boldsymbol{x}_{r1,G}-\boldsymbol{x}_{r2,G})$
9	*DE/best/2/bin*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{best,G}+F\cdot(\boldsymbol{x}_{r1,G}+\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G}-\boldsymbol{x}_{r4,G})$
10	*DE/rand/2/bin*	$\boldsymbol{v}_{i,G}=\boldsymbol{x}_{r5,G}+F\cdot(\boldsymbol{x}_{r1,G}+\boldsymbol{x}_{r2,G}-\boldsymbol{x}_{r3,G}-\boldsymbol{x}_{r4,G})$

表中 $\boldsymbol{x}_{best,G}$ 表示第 G 代种群中最好的个体；r_i 为随机整数，表示个体在种群中的序号。

$$f(x)=\frac{|S^x|^2}{\sum_{f_i,f_j\in S^x}C_{ij}^V} \tag{4}$$

其中：$C_{ij}^V=\dfrac{\sum_{(x,y)\in V}(f_i(x)-y)(f_j(x)-y)}{|V|}$ 表示神经网络 f_i，f_j 在验证集上的相关度。

微分进化算法中变异向量 $\boldsymbol{V}$ 及 $\boldsymbol{X}$ 都采用二进制，每次操作后对向量中的元素 x_i 先取其小数位值，如果小数位的值绝对值大于 0.5 则 x_i=1，否则 x_i=0。例如：

x=(–1.3,–6.8,1.3,6.8)→(–0.3,–0.8,0.3,0.8)→(0,1,0,1)

1.3 广义选择性神经网络集成

传统的选择性神经网络集成，只做一次选择，只生成一个神经网络集成，由于当已训练神经网络集成的个数 N 比较大（N>30）时，求得最优解几乎是不现实的，一般是求得近似最优解，基于此，笔者提出由不同的微分进化算法得到不同的近似解，同时得到不同的神经网络集成，然后从一系列不同的神经网络集成中选取一些性能优良的集成组成最后的神经网络集成，第一层生成神经网络集成时采取并行策略，时间复杂性影响不大，图 1 是广义选择性神经网络集成（Generalized Selective Neural Network Ensemble，GSNNE）的系统结构。

2 基于广义选择性神经网络集成的 MHC–II 分子结合肽的预测

2.1 数据预处理

MHC 分子结合肽预测分为 I、II 两类，其中 MHC I 类分子结合肽都有严格的长度[11]，通常由 9 个氨基酸残基组成，MHC

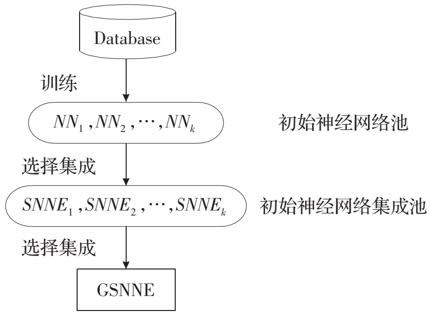

图 1　GSNNE 的系统结构

NN_i 表示第 i 个神经网络,$SNNE_i$ 表示第 i 个神经网络集成

II 类分子结合肽的核心区约 13 个氨基酸残基，但仅含有 9 个氨基酸的肽段为与 MHC II 类分子结合所必需,其中一个主要的锚定残基为结合必需，其他数个次级锚定残基影响结合,在已知相关等位基因的相关锚点知识以及根据已知知识所获得的对应该等位基因的联配矩阵的基础上,判断多肽与 MHC 分子的结合能力,分为低、中、高亲和力和非结合肽 4 大类,所以 MHC II 类分子结合肽通常为 4 分类问题或 2 分类问题(结合肽和非结合肽),预处理过程主要是:首先是数据抽取,来自一些公开的数据库,然后用联配矩阵以及相关的锚点知识将不等长的肽序变为等长的肽序(9),然后利用已有的联配矩阵(或用相关算法计算)把肽分为 4 类:高结合、中结合、低结合和不结合肽(或者分为结合肽和不结合肽两类),最后把数据分为训练集,测试集和验证集。

2.2　神经网络训练

神经网络集成中各个个体网络的差异性主要集中于:训练算法,网络的拓扑结构,隐含层神经元的个数和初始权重的设置等等,在神经网络个体生成后,利用已知的九肽训练样本对神经网络进行训练,对每一个氨基酸用 20 为二进制编码,具体如表 1,每一个九聚肽用 180 为二进制编码,比如 VYFYLQW-GR 编码为：000000000000000001000000000000000000000100001000000000000000000000000000000000001000000000100000000000000000000100000000000000000000000001000000100000000000000000000000000100000。

2.3　系统结构

步骤 1 数据抽取:数据来自不同的数据库 ,包括公开的或私人的数据库。

步骤 2 生成一系列初始选择性神经网络集成，然后从一系列初始选择性神经网络中选择一部分组成最后的神经网络集成,即广义选择性神经网络集成。

步骤 3 用广义选择性神经网络集成预测 MHC-II 类分子结肽(图 2 是 MHC 结合肽预测流程图)。

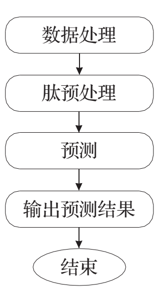

图 2　MHC 结合肽预测流程图

3　实验与分析

实验数据来自于标准数据库 MHCBench,该数据库主要是评价 MHC-II 分子结合预测方法的主要数据库，本文用的是 MHCBench 中的 SET-IV,该子集包含相同数目的 HLA-DRB1 *0401 结合肽(323)和非结合肽(323),多余和重复的肽被踢除。近似等可能的分成 6 个子集,每个子集中包含等可能多的结合和非结合肽,其中一个作为测试集,4 个作为训练集,余下的作为验证集。

本文实验的主要目的是与传统的选择性神经网络集成比较,为了简单起见,本文的 MHC II 类分子预测采用是是二分类模式(分类为结合肽和非结合肽两种),在初始神经网络池生成时,采用常用的不同类型的反向传播算法,隐含层神经元数有一定的区别,输入层包含 180 个节点,输出层神经元数为 1,采取并行策略训练出 21 个神经网络个体，第一层初始选择性神经网络集成池中包含有 9 个选择性神经网络集成($SNNE_i$,i=1,…,9),初始神经网络集成生成时采用 3 种不同的微分进化算法(DE_i,i=1,2,3),它们分别是:

$$DE_1: V_i^K = X_{r1}^K + F \cdot (X_{r2}^K - X_{r3}^K) \tag{5}$$

$$DE_2: V_i^K = X_i^K + F \cdot (X_{best}^K - X_i^K) + F \cdot (X_{r1}^K - X_{r2}^K) \tag{6}$$

$$DE_3: V_i^K = X_{best}^K + F \cdot (X_{r1}^K - X_{r2}^K) + F \cdot (X_{r3}^K - X_{r4}^K) \tag{7}$$

最终的广义选择性神经网络集成(GSNNE)是由第一层中 5 个预测精度高的 SNNE 集成。然后用 GSNNE 进行预测,预测实验重复 10 次,取平均值进行比较,表 2 是实验结果,表 2 的数据表明本文的广义神经网络集成优于传统的神经网络集成,从实验上说明了本文新范式的可行性和有效性。

表 2　广义选择性神经网络集成与选择性神经网络集成的预测结果　　%

模型	预测精度	
	非结合肽	结合肽
广义选择性神经网络集成(GSNNE)	99.16	98.42
选择性神经网络集成(SNNE)	91.97	90.15

SNNE 的预测结果是 5 个参与最后广义选择性神经网络集成中的最优值。

4　结论

为了进一步研究神经网络集成,本文提出了一种新的选择性神经网络集成范式，可以说是传统选择性神经网络的推广,实验表明该模型优于传统的选择性神经网络集成。然而该模型仅仅是一种新的思想,还没有严格的理论分析,这是未来值得探讨的问题,另外既然是一种集成模型,那么最后怎么集成以及初始选择性神经网络集成怎么生成也是非常有意义的研究课题,目前的仅仅是在分类方面的尝试,在回归问题上的应用也是很值得去探寻的,这些也是笔者所在的研究团队未来努力的方向。

参考文献：

[1] Buus S.Escription and prediction of peptide-MHC binding:the human MHC project[J].Curr Opin Immunol,1999,11:209-213.

[2] Singh H,Raghava G P S.ProPred:prediction of HLA-DR binding sites[J].Bioinformatics,2001,17:1236-1237.

(下转 63 页)