

基于免疫克隆算法的 LVQ 聚类算法权值优化

张晓丹 黄海燕

(华东理工大学信息科学与工程学院 上海 200237)

摘 要 学习矢量量化(LVQ)聚类算法存在严重的对初值敏感的问题,若初值的选择偏差太大,就不会产生好的聚类效果,致使聚类精度不够。免疫克隆算法具有很强的群体搜索能力,将免疫克隆算法用于优化 LVQ 聚类算法的初值,并将改进得到的聚类算法用于对 IRIS 数据集进行分类。分类结果与标准的 LVQ 算法的比较表明,改进后的聚类算法在稳定性上有了较大幅度的提高。

关键词 LVQ 聚类算法,免疫克隆算法,优化

中图法分类号 TP18 文献标识码 A

Optimization of Weights of LVQ Clustering Algorithm Based on Immune Clonal Algorithm

ZHANG Xiao-dan HUANG Hai-yan

(School Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract A major problem of learning vector quantization(LVQ) clustering algorithm is its sensitivity to the initialization, affecting the clustering precision. This paper first introduced the theory of LVQ clustering algorithm, then used immune clonal algorithm to optimize the initial values of LVQ. And the paper used IRIS data to test this new method. Comparing the evolving LVQ with the standard LVQ, the experiment results indicate that the approach of LVQ based on immune clonal algorithm has obvious stability to initial weights.

Keywords LVQ clustering algorithm, Immune clonal algorithm, Optimization

1 引言

LVQ(Learning Vector Quantization)算法是 Kohonen 提出的一种基于统计分布的自适应数据分类算法,它结合了监督与非监督的学习思想,在学习速率上较其它学习算法有着很大的提高,故将其应用于分类工作上,颇为省时方便。但是, LVQ 算法基于胜者为王的思想,存在神经元未被充分利用以及算法对初值敏感的问题,若初值的选择偏差太大,就不会产生好的聚类效果。免疫克隆算法^[2](Immune Clonal Algorithm, ICA)作为一种新的全局优化搜索算法,是模拟自然免疫系统功能的一种新的智能方法。它在传统进化算法的基础上,引入了亲和度和克隆机理,以保证算法能快速地收敛到全局最优解,成为近 40 年来免疫学中占主导地位的理论。

本文采用免疫克隆算法对 LVQ 聚类算法的初始权值进行优化,形成一种新的学习方法(简称为 ICALVQ 算法)。将 ICALVQ 算法用于对 UCI 数据集中的 IRIS 数据进行分类,结果表明,改进后的聚类算法在稳定性上较标准的 LVQ 聚类算法有了大幅度的提高,同时也意味着,基于免疫克隆算法的 LVQ 聚类算法有着良好的应用前景。

2 LVQ 聚类算法

LVQ 算法的网络模型如图 1 所示,由输入层、竞争层和输出层 3 部分组成。竞争层用于对输入向量进行分类,输出

层将竞争层传递过来的分类信息转换为使用者所定义的期望类别。在 LVQ 算法学习过程中,输入层与竞争层间的连接值逐渐被调整为聚类中心。

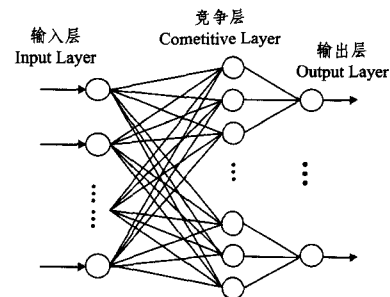


图 1 LVQ 聚类算法网络模型

LVQ 聚类算法模型的数学描述如下:

设输入向量用 X 表示:

$$X = (x_1 \ x_2 \ \cdots \ x_n)^T$$

竞争层的输出用 Y 表示:

$$Y = (y_1 \ y_2 \ \cdots \ y_m)^T$$

$$y_j \in \{0 \ 1\}, j=1, 2, \dots, m.$$

输出层的输出用 O 表示:

$$O = (o_1 \ o_2 \ \cdots \ o_l)^T$$

网络的期望输出用 d 表示:

$$d = (d_1 \ d_2 \ \cdots \ d_l)^T$$

输入层到竞争层之间的权值矩阵用 W^1 表示:

$$W^1 = (W_1^1 \ W_2^1 \ \cdots \ W_j^1 \ \cdots \ W_m^1)$$

其中,列向量 W_j^1 为竞争层第 j 个神经元对应的权值向量。

竞争层到输出层之间的权值矩阵用 W^2 表示:

$$W^2 = (W_1^2 \ W_2^2 \ \cdots \ W_k^2 \ \cdots \ W_l^2)$$

其中,列向量 W_k^2 为输出层第 k 个神经元对应的权值向量。 n 为输入层神经元个数,取决于输入向量的维数。 m 为竞争层神经元的个数,也是聚类中心的个数。 l 为输出层神经元个数,由期望类别决定。网络学习速率为 η ,训练次数为 t_m 。

LVQ 聚类算法的学习步骤如下:

Step1 初始化。

对权值 W^1, W^2 进行初始设定:随机取较小的 W^1 的行向量,赋值后的 $W^1(0)$ 即为 LVQ 神经网络的初始权值。对于 W^2 ,当某个竞争层的神经元权值向量对应的子类属于线性层某个神经元所对应的期望类别时,这两个神经元之间的连接权值设置为 1,否则设置为 0。 W^2 一旦设定,就不再改变。确定初始学习速率 $\eta(0) \in (0 \ 1)$ 和最大训练次数 t_m 。

Step2 输入样本向量 X 。

计算此时输入样本向量 $X(t)$ 与表示各子类聚类中心的 W^1 各行向量的欧式距离:

$$D = \begin{bmatrix} \|W_1^1 - X(t)\| \\ \|W_2^1 - X(t)\| \\ \vdots \\ \|W_m^1 - X(t)\| \end{bmatrix}$$

Step3 寻找获胜神经元 j^* 。

找出距离最近的权值行向量来确定获胜的竞争层神经元 j^* :

$$\|X - W_{j^*}^1\| = \min_j \|X - W_j^1\|$$

把竞争层的输出向量 Y 的第 j^* 个元素设定为 1,其它为 0。

Step4 计算线性输出层输出向量 O 的值。

$O = W^2 Y$, O 中唯一的一个非零元素其序号 k^* 就表明 $X(t)$ 分为第 k^* 类。

Step5 比较网络实际输出 O 与目标期望输出 d 。

若分类正确,按照下式修正 $W_{j^*}^1$:

$$W_{j^*}^1(t+1) = W_{j^*}^1(t) + \eta(t)[X - W_{j^*}^1(t)]$$

否则逆输入样本方向调整权值:

$$W_{j^*}^1(t+1) = W_{j^*}^1(t) - \eta(t)[X - W_{j^*}^1(t)]$$

其他非获胜神经元的权值保持不变。

Step6 更新学习速率。

$$\eta(t) = \eta(0)(1 - \frac{t}{t_m})$$

当 $t < t_m$ 时, $t = t + 1$, 转到 Step2 输入下一个样本,重复各步骤直到 $t = t_m$ 。

3 结合免疫克隆算法的 LVQ 算法

结合 LVQ 聚类算法的学习过程,若 LVQ 算法的初始权值 $W^1(0)$ 赋值不当(即与理想的聚类中心偏离较大),那么神经元在网络训练过程中成为获胜神经元 j^* 的几率较小,其对应的权值 $W_{j^*}^1$ 得到调整的几率也相应降低,网络训练结束后,输入层到竞争层间的连接值与理想值间的差距也会越大,分类效果则不明显,可见 LVQ 算法对初值的敏感程度较高。

免疫克隆算法具有很强的群体搜索能力,可以用来为 LVQ 聚类算法选择合适的初值。

3.1 免疫克隆算法

免疫克隆算法的核心是克隆算子,克隆算子具有很强的群体搜索能力。该算子主要包含 3 种操作:克隆操作、免疫基因操作和克隆选择操作,其基本流程如下:

Step1 $k=0$, 初始化抗体种群 $A(0)$, 设定算法参数, 计算初始种群的亲和度;

Step2 停止准则判断,若符合停止规则,算法停止,否则对抗体种群 $A(k)$ 进行克隆操作,得到克隆后的种群 $Y(k)$;

Step3 以概率 P_m 对 $Y(k)$ 实现免疫基因操作,得到新的抗体种群 $Z(k)$;

Step4 计算 $A'(k) = Z(k) \cup A(k)$ 的亲和度;

Step5 克隆选择操作:从 $A'(k)$ 中选出 n 个亲和度较大的抗体,形成新的抗体种群 $A(k+1)$, $k=k+1$, 返回 Step2。

免疫克隆算法的群演化过程如图 2 所示。其中克隆操作有效地实现了搜索空间的扩张;免疫基因操作实现了抗体亲和度的成熟和多样性的产生;克隆选择操作通过局部择优,有效地压缩了种群的大小。

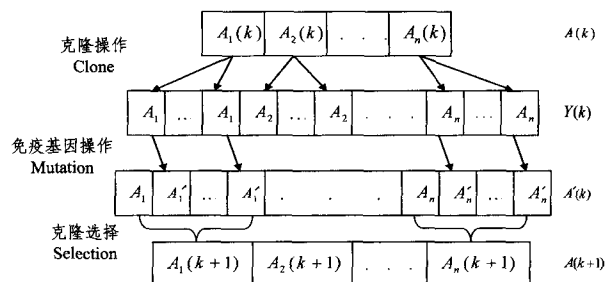


图2 克隆选择算法的主要操作过程

3.2 基于免疫克隆算法优化的 LVQ 算法

结合免疫克隆算法优化 LVQ 聚类算法的初值,首先通过免疫克隆算法(ICA)获得最佳的初值,然后应用 LVQ 聚类算法实现对输入向量的快速、准确识别,现步骤如下:

Step1 定义 ICA 参数。

随机产生 m 个 n 维向量 $v_{k,0} (k=1, 2, \dots, m)$ 作为最初的抗体 $A(0)$, $A(0) = [v_{1,0} \ \cdots \ v_{m,0}]$ 。确定克隆规模 q 、免疫基因变异概率 p_m 及误差参数 $\epsilon, t=0$ 。

Step2 计算初始种群的亲和度。

亲和度函数为 $f(t) = \frac{1}{1 + D_{k,t}}$, $D_{k,t}$ 是个体到抗体的平均

距离, $D_{k,t} = \frac{1}{N} \sum_j \|x_j - v_{k,t}\|$, x_j 表示输入样本向量, N 表示群体数目。

Step3 通过 ICA 确定初始权值。

①判断是否满足终止条件:

$d = \sum_{k=1}^m D_{k,t} - \sum_{k=1}^m D_{k,t-1}$, 若 $|d| < \epsilon$, 则迭代终止, 确定当前种群为最佳初始权值, 否则继续。

②依据克隆规模 q 对抗体进行克隆操作。

③依据概率 p_m 对克隆后的抗体进行免疫基因操作。

④依据免疫操作后抗体亲和力的大小, 进行克隆选择操作。

⑤ $k=k+1$, 返回①。

(下转第 53 页)

[3] 疫情动态[J]. 中国动物保健, 2012(5)

[4] 王丽萍, 徐友富, 王建军, 等. 1990—2006 年安徽疟疾疫情时空分布特点研究[J]. 疾病控制杂志, 2008, 12(2)

[5] 叶敏, 李晓松, 殷菲. 2004 年我国麻疹发病情况分析[J]. 现代预防医学, 2008, 35(8)

[6] 武继磊, 王劲峰, 孟斌, 等. 2003 年北京市 SARS 疫情空间相关性分析[J]. 浙江大学学报, 2005, 31(1)

[7] MacIntosh V H. Enhancing Influenza Surveillance Using Electronic Surveillance System for the Early Notification of Community-Based Epidemics[C]//NTIS. 2004: 1-17

[8] Xiang Ji, Chun S A, Geller J. Epidemic outbreak and spread detection system based on twitter data[C]//ICHIS. 2012: 152-163

[9] 赵红蕊, 唐中实. 基于图像空间聚类的滤波技术[J]. 计算机应用, 2006, 26(11)

[10] 王海军, 张德礼. 基于空间聚类的城镇土地定级方法研究[J]. 武汉大学学报, 2006, 31(7)

[11] 梅新, 崔伟宏, 高飞, 等. 基于空间聚类的物流配送决策研究[J]. 武汉大学学报, 2008, 33(4)

[12] Figuera C, Lillo J M, Mora-Jiménez I, et al. Multivariate spatial clustering of traffic accidents for local profiling of risk factors [C]//14th International IEEE Conference on Intelligent Transportation Systems(ITSC). 2011

[13] Cornelis B. Framing Spatial Decision-Making and Disaster Management in Time, Geo-information for Disaster Management [M]. Springer, 2005, Part 3: 281-293

[14] 郭茂耘. 航天发射安全控制决策的空间信息分析与处理研究 [D]. 重庆: 重庆大学, 2011

[15] 黄先锋, 程晓光, 张帆, 等. 基于边长比约束的离散点准确边界追踪算法[J]. 武汉大学学报, 2009, 34(6)

[16] 袁满, 袁志华. 一种基于行列法离散点边界搜索算法[J]. 计算机应用研究, 2010, 27(1)

(上接第 28 页)

Step4 将获得的初值赋给 LVQ 算法。

利用免疫克隆算法对初值进行选择, 将选择结果赋值给 LVQ 作为其初始权值。论文结合免疫克隆聚类算法对 LVQ 聚类算法进行改进, 改进后的算法简称为 ICALVQ 算法。ICALVQ 算法的流程图如图 3 所示。

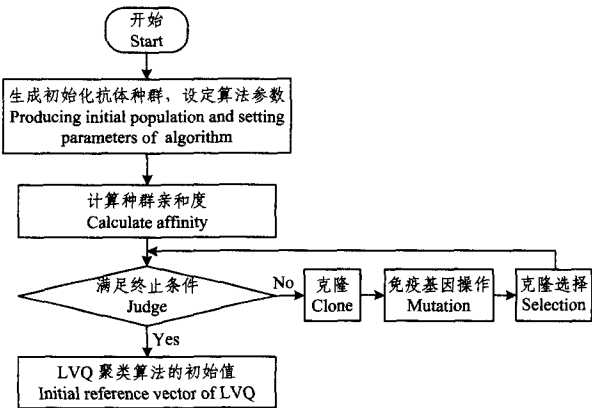


图 3 ICALVQ 算法流程图

4 实验结果

为了验证 ICALVQ 算法的有效性, 本文选择 UCI 数据集的 IRIS 数据进行测试。IRIS 数据集包含 150 个数据组, 分为 3 类, 每类 50 个数据, 每个数据包含 4 个属性, 是在数据挖掘、数据分类中非常常用的测试集、训练集。

设定免疫克隆算法的克隆规模 q 为定值 5, LVQ 算法的学习次数为 500。分别采用文献[5]中的初始权值(3 个聚类中心的值均为 0 或 9)以及接近聚类中心的初始权值进行试验, 测试结果分别见表 1 和表 2。

由表 1 表 2 看出, 当选择的初值远离聚类中心时, 传统 LVQ 算法学习过程中只有 v_1 得到了调整, v_2 和 v_3 由于未成为获胜神经元而始终没有得到调整, 只有当选择的初值接近聚类中心时, 传统的 LVQ 最终学习后得到的聚类中心才较为合理。而 ICALVQ 算法则始终保持稳定性, 很好地克服了 LVQ 算法所存在的对初值敏感的缺点。

表 1 初始权值远离聚类中心时两种算法最终结果的比较

初值	LVQ 算法的聚类中心	ICALVQ 算法的聚类中心
$v_1 = (0 \ 0 \ 0 \ 0)$	$v_1 = (5.836 \ 3.057 \ 3.752 \ 1.201)$	$v_1 = (6.599 \ 2.985 \ 5.564 \ 2.037)$
$v_2 = (0 \ 0 \ 0 \ 0)$	$v_2 = (0.0 \ 0.0 \ 0.0 \ 0.0)$	$v_2 = (5.926 \ 2.671 \ 4.351 \ 1.406)$
$v_3 = (0 \ 0 \ 0 \ 0)$	$v_3 = (0.0 \ 0.0 \ 0.0 \ 0.0)$	$v_3 = (5.006 \ 3.415 \ 1.472 \ 0.252)$
$v_1 = (9 \ 9 \ 9 \ 9)$	$v_1 = (5.836 \ 3.057 \ 3.752 \ 1.201)$	$v_1 = (6.599 \ 2.985 \ 5.564 \ 2.037)$
$v_2 = (9 \ 9 \ 9 \ 9)$	$v_2 = (9.0 \ 9.0 \ 9.0 \ 9.0)$	$v_2 = (5.926 \ 2.671 \ 4.351 \ 1.406)$
$v_3 = (9 \ 9 \ 9 \ 9)$	$v_3 = (9.0 \ 9.0 \ 9.0 \ 9.0)$	$v_3 = (5.006 \ 3.415 \ 1.472 \ 0.252)$

表 2 初始权值接近聚类中心时两种算法最终结果的比较

初值	LVQ 算法的聚类中心	ICALVQ 算法的聚类中心
	$v_1 = (6.846 \ 3.076 \ 5.745 \ 2.076)$	$v_1 = (6.599 \ 2.985 \ 5.564 \ 2.037)$
$v_1 = (6 \ 3 \ 5 \ 2)$	$v_2 = (5.896 \ 2.732 \ 4.431 \ 1.447)$	$v_2 = (5.926 \ 2.671 \ 4.351 \ 1.406)$
$v_2 = (6 \ 3 \ 4 \ 1)$	$v_3 = (5.004 \ 3.426 \ 1.461 \ 0.246)$	$v_3 = (5.006 \ 3.415 \ 1.472 \ 0.252)$
$v_3 = (4 \ 3 \ 1 \ 0.1)$		

结束语 论文首先对 LVQ 算法进行了深入的分析, 基于免疫克隆算法具有较强群体搜索能力的特点, 将免疫克隆算法运用到 LVQ 算法的优化问题中, 以克服传统 LVQ 算法对初始权值敏感的缺点。将基于免疫克隆算法的 LVQ 新算法应用到 IRIS 数据集的分类实验上, 实验结果表明, ICALVQ 算法较传统 LVQ 算法有着很强的稳定性, 从而有效地克服了 LVQ 算法对初始权值敏感的问题, 显示出其良好的应用前景。

参考文献

[1] Jiao L C, Wang L. A novel genetic algorithm based on Immunity [J]. IEEE Transaction on Systems, Man, and Cybernetics-part A: Systems and Humans, 2000, 30(5): 552-551

[2] 焦李成, 杜海峰, 刘芳, 等. 免疫优化计算、学习与识别[M]. 北京: 科学出版社, 2006: 92-99

[3] 马文萍, 尚荣华, 焦李成. 免疫克隆优化聚类技术[J]. 西安电子科技大学学报: 自然科学版, 2007, 34(6): 911-918

[4] 张敏灵, 陈兆乾, 周志华. SOM 算法、LVQ 算法及其变体综述 [J]. 计算机科学, 2002, 29(7): 97-100

[5] Pal N R, Bezdek J C, Tsao E C K. Generalized clustering networks and Kohonen's Self-organizing scheme[J]. IEEE Transactions on Neural Networks, 1993, 4(4): 549-557