第 30 卷 第 4 期
2004 年 7 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 30, No. 4
July, 2004

MLVQ 网络聚类算法[1)]

闫德勤[1] 迟忠先[2] 王 军[3]

[1](辽宁师范大学计算机系 大连 116029)

[2](大连理工大学计算机系 大连 116024)

[3](大连理工大学应用数学系 大连 116024)

(E-mail: yandeqin@163.com)

摘 要 讨论了关于改进 LVQ 聚类网络的理论与算法. 为克服 LVQ 网络聚类算法对初值敏感的问题广义学习矢量量化(GLVQ)网络算法对 LVQ 算法进行了改进,但 GLVQ 算法性能不稳定. GLVQ-F 是对 GLVQ 网络算法的修改,但 GLVQ-F 算法仍存在对初值的敏感问题. 分析了 GLVQ-F 网络算法对初值敏感的原因以及算法不稳定的理论缺陷,改进了算法理论并给出了一种新的改进的网络算法(MLVQ). 实验结果表明新的算法解决了原有算法所存在的问题,而且性能稳定.

关键词 聚类分析,GLVQ-F 算法,MLVQ 算法

中图分类号 TP301

MLVQ Clustering Neural Network

YAN De-Qin[1] CHI Zhong-Xian[2] WANG Jun[3]

[1](*Department of Computer Science, Liaoning Normal University, Dalian* 116029)

[2](*Department of Computer Science, Dalian University of Technology, Dalian* 116024)

[3](*Department of Applied Mathematics, Dalian University of Technology, Dalian* 116024)

(E-mail: yandeqin@163.com)

Abstract Theory and algorithms of clustering neural network are discussed. The generalized LVQ neural network was aimed to improve the algorithm of LVQ. However it brought another problem. GLVQ-F algorithm was an improvement to GLVQ, but it was still unstable and left the problem of LVQ unsolved. The defect of GLVQ-F clustering neural network algorithm is theoretically analyzed in this paper, and the modified theory is discussed. Finally, a modified neural network algorithm of competitive learning schemes (MLVQ) is given. Experiment results show the new algorithm is stable and effective.

Key words LVQ algorithm, cluster analysis, MLVQ algorithm

1) 国家自然科学基金(19831050)资助
Supported by National Natural Science Foundation of P. R. China(19831050)
收稿日期 2002-04-08 收修改稿日期 2002-12-16
Received April 8, 2002; in revised form December 16, 2002

1　引言

Kohonen[1]的学习矢量量化(LVQ)网络算法为聚类分析提供了一个新的思路，但该算法的一个主要问题是对初值的敏感性[2]. 为了解决学习矢量量化(LVQ)网络算法所出现的问题，Pal 等[2]提出了一种广义学习矢量量化(GLVQ)网络算法. 然而，在原始数据成比例变化的情况下该网络失去了有效的聚类效果[3]. 因此，Karayiannis 提出了关于 GLVQ 的修正算法[4](GLVQ-F). GLVQ-F 算法在一定程度上解决了 GLVQ 算法存在的问题，但 GLVQ-F 算法仍存在对初值的敏感问题，而且当模糊度因子 m 变化时其算法不稳定[5]. 本文分析了 GLVQ-F 网络算法对初值的敏感的原因以及算法不稳定的理论缺陷，提出了一种新的改进算法 MLVQ. 实验结果表明新的算法解决了原有算法所存在的问题，而且性能稳定有效.

2　关于 LVQ 及 GLVQ 与 GLVQ-F 网络聚类算法

设 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\subset R^p$ 为 n 个样本，其聚类个数为 c，聚类中心(或称原型)为 $V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\}$. 为求得聚类中心，LVQ 采用两层神经网络：每个样本从输入层输入，通过一定的权值连到输出层的每一节点 $\boldsymbol{v}_i(i=1,2,\cdots,c)$. 通过更新获胜竞争节点的方法使网络不断更新直到收敛. 更新节点(又称网络的学习)的方法为

$$\boldsymbol{v}_{i,t}=\boldsymbol{v}_{i,t-1}+\alpha_t(\boldsymbol{x}_k-\boldsymbol{v}_{i,t-1}) \tag{1}$$

其中 $\alpha_t=\alpha_{t-1}\left(1-\frac{t}{T}\right)$ 为学习因子，t 为学习次数，T 为学习总次数.

由于 LVQ 网络聚类算法对初值敏感，Pal 等在文献[2]中在 LVQ 算法的基础上做了改进：给出亏损函数 L_x 作为目标函数，对竞争元优化目标函数以确定学习因子(诸竞争元的学习方向为其对目标函数的负梯度方向). 文献[2]给出的亏损函数为 $L_x=\sum_{r=1}^{c}g_r\|\boldsymbol{x}-\boldsymbol{v}_r\|^2$，其中 g_r 是样本 $\boldsymbol{x}$ 相对于原型 $\boldsymbol{v}_r$ 的亏损因子.

由于 GLVQ 算法对样本数据按比例变化后聚类出现问题，文献[4]文讨论了 g_r 的选取方法，得出 GLVQ 的修改算法(GLVQ-F 算法). GLVQ-F 算法的学习方式为

$$\boldsymbol{v}_{i,t}=\boldsymbol{v}_{i,t-1}+\left(\frac{2\alpha_t}{m-1}\right)(\boldsymbol{x}_k-\boldsymbol{v}_{i,t-1})\left[(m-2)u_i+\left(\sum_{r=1}^{c}\left(\frac{\|\boldsymbol{x}-\boldsymbol{v}_i\|^{2/(m-1)}}{\|\boldsymbol{x}-\boldsymbol{v}_r\|^{2/(m-1)}}\right)\right)^{2-m}u_i^2\right] \tag{2}$$

其中 u_i 是样本点对聚类中心的隶属度，m 是模糊度常数因子.

但 GLVQ-F 算法存在着两个问题：一是对初值的敏感性(把 3 个聚类中心都聚在一点上)，另一个就是算法的不稳定(对无监督分类时，随 m 的变化分类结果不稳定).

3　GLVQ-F 网络算法的问题分析

GLVQ-F 的聚类网络采用的仍是同 LVQ 一样的两层神经网络，网络运行方式是每一个输出元都参加学习，学习方式见式(2). 比较 LVQ 竞争学习网络以及由式(2)可知，

GLVQ-F 网络的每一个输出元都相当于一个竞争获胜元. 当对聚类中心赋予同一初值时，对它们的更新没有区别. 从网络运行的过程可以看出这些输出元的值会变为一个，从而失去聚类的作用. 这正是 GLVQ-F 网络算法对初值敏感的原因.

亏损函数的亏损因子有多种选法，GLVQ-F 把 GLVQ 的亏损因子 g_r 用模糊均值(FCM)算法[5]中的隶属度 u_r 代替导出了网络学习因子的另一种形式. 这样就存在两个问题. 一个是对亏损函数的选取缺乏理论支持. 亏损函数实质的意义是能量函数，应根据能量表示去产生. 而模糊均值(FCM)算法中 u_r 本身关联着 m，能量函数的完整形式是

$$L_x = \sum_{r=1}^{c} u_r^m \| \boldsymbol{x} - \boldsymbol{v}_r \|^2 \tag{3}$$

因此，GLVQ-F 的亏损函数的选取存在理论问题. 另一个是 u_r 的使用不当. GLVQ-F 方法中把 u_r 用来代替 GLVQ 中亏损函数的 g_r 以求得网络的学习因子，导致算法不合理. 这两个问题是 GLVQ-F 算法不稳定的根源.

4 新的聚类网络算法(MLVQ)

通过对 GLVQ-F 的算法分析可以看到，为了设计合理的聚类神经网络必须：

a) 所有的竞争元都要参加学习；

b) 亏损函数与亏损因子的选择要有合理的理论依据.

为此，新的聚类改进方法仍然采用 LVQ 两层网络，算法如下：首先采用一种对初值的处理算法，这种算法既保证每个输出元机遇均等又使得各有差别；然后以 FCM 的目标函数为亏损函数，优化后求出学习因子；对每一竞争元依据不同的学习因子进行学习直到满足收敛条件. 我们称改进后的算法为 MLVQ 网络聚类算法. MLVQ 算法结构如下：

1) 选定误差 ε_1 或首次学习次数 T_1 (一般 $T_1 = 2c$)，把 LVQ 算法中的判获胜元方式 $\| \boldsymbol{x}_k - \boldsymbol{v}_{i,t-1} \| = \min\limits_{1 \leqslant j \leqslant c} \{ \| \boldsymbol{x}_k - \boldsymbol{v}_{j,t-1} \| \}$ 变为 $\| \boldsymbol{x}_k - \boldsymbol{v}_{i,t-1} \| = \max\limits_{1 \leqslant j \leqslant c} \{ \| \boldsymbol{x}_k - \boldsymbol{v}_{j,t-1} \| \}$，对获胜元进行学习

$$\boldsymbol{v}_{i,t} = \boldsymbol{v}_{i,t-1} + 2\alpha_t (\boldsymbol{x}_k - \boldsymbol{v}_{i,t-1}) u_i^m$$

直至满足选定的误差 ε_1 或首次学习次数 T_1；

2) 以 1)的结果为初值 $V_0 = \{\boldsymbol{v}_{1,0}, \boldsymbol{v}_{2,0}, \cdots, \boldsymbol{v}_{c,0}\} \subset R^p$ 进行如下计算，设定最大循环次数 T 以及误差 ε，顺次选定聚类竞争元进行学习

$$\boldsymbol{v}_{i,t} = \boldsymbol{v}_{i,t-1} + 2\alpha_t (\boldsymbol{x}_k - \boldsymbol{v}_{i,t-1}) u_i^m \tag{4}$$

直到满足收敛条件或最大循环次数 T，其中

$$u_i = \left(\sum_{j=1}^{c} \left(\frac{\| \boldsymbol{x} - \boldsymbol{v}_i \|^{2/(m-1)}}{\| \boldsymbol{x} - \boldsymbol{v}_j \|^{2/(m-1)}} \right) \right)^{-1}$$

式(4)中学习因子 $2\alpha_t u_i^m$ 的得出参看附录 A.

5 实验结果

为便于比较，我们仍用著名的 Iris 数据样本对新算法进行实验，并采用文献[4]的实验条件：α_1 与学习总次数 T 分别为 0.6 和 500，$m=2$. 表 1 列出了对于三个聚类中心的初值都

是 0 或 9(文献[2]所用的初值)时分别用 GLVQ-F 和 MLVQ 算法所算出的聚类中心. 从表中可以清楚地看出 GLVQ-F 算法对于这样的初值类似 LVQ 算法：对初值的敏感. 而 MLVQ 算法则较好地解决了这个问题. 表 2 列出的是文献[4]的计算条件，用 Iris 数据以及除以 10 以后的 Iris 数据分别用 GLVQ-F 和 MLVQ 算法所算出的聚类中心. 可见，MLVQ 算法较好地解决了 GLVQ-F 算法所要解决的问题.

表 1　GLVQ-F 与 MLVQ 算法在不同初值时对 Iris 数据的聚类

Table 1　The clustering results cn Iris data by the GLVQ-F and the MLVQ algorithms for different initializatioins

初值	GLVQ-F 算法的聚类中心	MLVQ 算法的聚类中心
$v_1(0,0,0,0)$	$v_1(5.884,3.039,3.874,1.246)$	$v_1(6.784,3.059,5.669,2.069)$
$v_2(0,0,0,0)$	$v_2(5.884,3.039,3.874,1.246)$	$v_2(5.901,2.754,4.427,1.421)$
$v_3(0,0,0,0)$	$v_3(5.884,3.039,3.874,1.246)$	$v_3(5.002,3.409,1.487,0.257)$
$v_1(9,9,9,9)$	$v_1(5.884,3.039,3.874,1.246)$	$v_1(6.784,3.059,5.669,2.069)$
$v_2(9,9,9,9)$	$v_2(5.884,3.039,3.874,1.246)$	$v_2(5.901,2.754,4.427,1.421)$
$v_3(9,9,9,9)$	$v_3(5.884,3.039,3.874,1.246)$	$v_3(5.002,3.409,1.487,0.257)$

表 2　GLVQ-F 与 MLVQ 算法对 Iris 及 Iris/10 数据的聚类

Table 2　The clustering results on Iris and Iris/10 data by the GLVQ-F and the MLVQ algorithms

初值	GLVQ-F 算法的聚类中心	MLVQ 算法的聚类中心
$v_1(7.9,4.4,6.9,2.5)$	$v_1(6.806,3.067,5.692,2.078)$	$v_1(6.784,3.059,5.669,2.069)$
$v_2(6.1,3.2,3.95,1.3)$	$v_2(5.917,2.758,4.459,1.439)$	$v_2(5.901,2.754,4.427,1.421)$
$v_3(4.3,2.0,1.0,0.1)$	$v_3(5.003,3.408,1.490,0.258)$	$v_3(5.002,3.409,1.487,0.257)$
$v_1(0.79,0.44,0.695,0.25)$	$v_1(0.680,0.306,0.569,0.2079)$	$v_1(0.6784,0.3059,0.5669,0.2069)$
$v_2(0.61,0.32,0.39,0.13)$	$v_2(0.591,0.275,0.445,0.1439)$	$v_2(0.5901,0.2754,0.4427,0.1421)$
$v_3(0.43,0.20,0.1,0.01)$	$v_3(0.500,0.340,0.149,0.0258)$	$v_3(0.5002,0.3409,0.1487,0.0257)$

6　新算法与 GLVQ-F 算法的比较

MLVG 算法与文献[4]的 GLVQ-F 算法有两个方面不同：1) 新算法 MLVQ 在分析了 LVQ 及 GLVQ-F 网络运行机理的基础上给出了有效的抗初值敏感方法，GLVQ-F 算法对所给初值都相同时出现聚类错误；2) 新算法 MLVQ 从亏损函数的实质意义出发使用完整的模糊均值(FCM)方法的目标函数作为亏损函数，从而导出网络的学习因子，GLVQ-F 把 GLVQ 的亏损因子 g_r 用模糊均值(FCM)方法中的模糊因子 u_r 代替导出了网络学习因子在数理关系上是错误的.

References

1　Kohonen T. Self-Organization Maps. Berlin: Springer-Verlag, 1995

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

3　Gonzalez A I, Grana M, Anjou A D. An analysis of the GLVQ algorithm. *IEEE Transactions on Neural Networks*, 1995, **6**(5): 1012～1016

4　Karayiannis N B, Bezdek J C, Pal N R, Hatharay R J. Repair to GLVQ: A new family of competitive learning schemes. *IEEE Transactions on Neural Networks*, 1996, **7**(5): 1062～1071

5　Bezdek J. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum, 1981

6　Zhang Zhi-Hua, Zheng Nan-Ning, Wang Tian-Shu. Behavioral analysis and improvement of generalized LVQ neural network. *Acta Automatica Sinica*, 1999, **25**(5): 583～589 (in Chinese)

附录 A

亏损函数定义为 FCM 的目标函数 $L_x=\sum_{r=1}^{c}u_r^m\|x-v_r\|^2$，令

$$\omega_i=u_i^{-1}=\sum_{j=1}^{c}\left(\frac{\|x-v_i\|^{2/(m-1)}}{\|x-v_j\|^{2/(m-1)}}\right),\quad i=1,2,\cdots,c$$

当 $r\neq k$ 时，

$$\frac{\partial}{\partial v_k}\omega_r=\|x-v_r\|^{2/(m-1)}\frac{\partial}{\partial v_k}(\|x-v_k\|^2)^{-1/(m-1)}=\frac{2}{m-1}(x-v_k)\|x-v_k\|^{-2}\left(\frac{\|x-v_r\|}{\|x-v_k\|}\right)^{2/(m-1)}\tag{A1}$$

当 $r=k$ 时，

$$\frac{\partial}{\partial v_k}\omega_k=\left(\frac{\partial}{\partial v_k}\|x-v_k\|^{2/(m-1)}\right)\sum_{j=1}^{c}\|x-v_j\|^{-2/(m-1)}=\frac{-2}{m-1}(x-v_k)\|x-v_k\|^{-2}(u_k^{-1}-1)\tag{A2}$$

由式(A1)和式(A2)，

$$\frac{\partial}{\partial v_k}u_r^m=-\frac{2m}{m-1}(x-v_k)u_r^{m+1}\|x-v_k\|^{-2}\left(\frac{\|x-v_r\|}{\|x-v_k\|}\right)^{2/(m-1)}$$

$$\frac{\partial}{\partial v_k}u_k^m=\frac{2m}{m-1}(x-v_k)\|x-v_k\|^{-2}(u_k^{-1}-1)u_k^{m+1}$$

所以，

$$\frac{\partial}{\partial v_k}L_x=\frac{2m}{m-1}(x-v_k)(u_k^{-1}-1)u_k^{m+1}-2(x-v_k)u_k^m-\frac{2m}{m-1}(x-v_k)\sum_{\substack{r=1\\r\neq k}}^{c}u_r^{m+1}\left(\frac{\|x-v_r\|}{\|x-v_k\|}\right)^{2m/(m-1)}$$

从而得到

$$\frac{\partial}{\partial v_k}L_x=-\frac{2m}{m-1}(x-v_k)u_k^m\left[\sum_{r=1}^{c}\left(\frac{\|x-v_k\|}{\|x-v_r\|}\right)^{2/(m-1)}u_k-\frac{1}{m}\right]=$$

$$-\frac{2m}{m-1}(x-v_k)u_k^m\left[1-\frac{1}{m}\right]=-2(x-v_k)u_k^m$$

闫德勤 教授，博士. 主要研究领域为模式识别、图像处理.

(**YAN De-Qin** Professor at Department of Computer Science, Liaoning Normal University. He received his Ph. D. degree at the Institute of Mathematics, Nankai University in 1999. His research interests include pattern recognition and image processing.)

迟忠先 教授，博士生导师. 主要研究领域为知识发现、数据仓库、数据挖掘等.

(**CHI Zhong-Xian** Professor and doctoral supervisor at Department of Computer Science, Dalian University of Technology. His research interests include database, data warehousing, knowledge discovery, and data mining.)

王 军 教授，博士，博士生导师. 主要研究领域为组合数学及图论.

(**WANG Jun** Professor and doctoral supervisor at Department of Computer Science, Dalian University of Technology. His research interests include combinatorics and combinatorial graph theory.)