混合 Neural-Gas 网络和 Sammon 映射的数据可视化算法

晋良念 欧阳缮
(桂林电子科技大学信息与通信学院 桂林 541004)

摘 要：与SOFM，最大熵聚类，K均值聚类相比，“Neural-Gas”网络算法具有收敛速度快、代价误差小等优点。但“Neural-Gas”网络用于非均匀分布的线性或非线性数据集进行降维或可视化时，输出空间上固定有序的神经元表现出极不理想的距离信息。为此，该文根据归一化概率自组织特征映射的基本思想，提出混合“Neural-Gas”网络和Sammon映射的新方法来解决此问题，通过“Neural-Gas”网络算法进行特征聚类以降低计算复杂度，通过Sammon映射保持输入空间和输出空间上神经元间的距离相似性。仿真结果表明，该混合算法对合成数据集或现实数据集的可视化能够取得较理想的效果，从而验证了该混合算法的可行性和有效性。

关键词：Neural-Gas网络；Sammon映射；混合算法；距离相似性
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2008)05-1118-04

Algorithm for Data Visualization by Hybridizing Neural Gas Network and Sammon's Mapping

Jin Liang-nian Ouyang Shan
(*School of Info and Commun.,Guilin University of Electron. Tech.*, *Guilin* 541004, *China*)

Abstract: Compared with Self-Organizing Feature Map(SOFM), maximum-entropy clustering and K-means clustering, the Neural-Gas network algorithm has advantages of faster convergence, smaller cost distortion errors, etc. However, the fixed and regular neurons on the output space represent worse distance information when the neural gas network algorithm is used for dimension reduction and visualization of linear or nonlinear data sets with nonuniform distribution. Therefore, according to the basic idea of the probabilistic regularized SOFM, a new visualization method for hybridizing neural gas network and Sammon's mapping is proposed to overcome this problem, and it reduces the computational complexity with using neural gas network algorithm for feature clustering and preserves the interneuronal distances resemblance from input space into output space by using Sammon's mapping. Simulation results show that the proposed hybridizing algorithm can obtain the better visualization effect on the synthetic and real data sets, thus demonstrating the feasibility and effectiveness of the hybridizing algorithm.

Key words: Neural-Gas network; Sammon's mapping; Hybridizing algorithm; Distances resemblance

1 引言

自组织特征映射(SOFM)、最大熵特征映射以及“Neural-Gas”网络具有聚类、降维、自学习以及可视化的功能，已广泛应用于模式识别、数据挖掘、故障诊断等领域[1–4]。这些方法导出的学习规则，对获胜神经元及邻域神经元的调整均以欧氏距离为度量，使它们能够在输出空间上保持输入数据空间的拓扑结构，并由此距离构成的分类器是一个垂直于两类超球体中心连线并通过连线中点的“超平面”，所以对样本分布类似超球体的数据集能取得较好的分类效果。与SOFM、最大熵聚类、K均值聚类相比，“Neural-Gas”网络的邻域核刻画了输入空间上数据点间的邻域信息，而非二维输出网格上点间的邻域信息，其学习算法具有收敛速度快、代价误差小及收敛性稳定等优点。所以该文重点针对“Neural-Gas”网络的可视化功能进行详细地分析。尽管该神经网络对类似超球体的均匀数据集能取到较好的聚类效果，并能保持空间拓扑关系，但现实中的诸多数据集分布呈现多态形式，具有高度非均匀性，甚至是线性不可分等特点，而“Neural-Gas”网络在固定有序的输出空间上不能自然地，真实地展现原输入空间的距离信息，因此其可视化效果极不理想。另外，传统的非线性降维方法，如多维量度(MDS)，Sammon映射[5, 6]尽管能保持映射空间与原空间的距离相似性，数据的可视化效果比较理想，但这些算法的计算复杂度较大，尤其对大数据集合的非线性映射几乎不可行。为此，该文应用归一化概率自组织特征映射的思想[7]，在深入分析“Neural-Gas”网络和Sammon映射的基础上，提出混合“Neural-Gas”网络和Sammon的新方法解决上述问题。

2006-10-13收到，2007-04-04改回
国家自然科学基金(60172011)资助课题

2　Neural-Gas 网络和 Sammon 非线性映射

2.1 Neural-Gas 网络[1, 2]

类似于最大熵聚类和SOFM，“Neural-Gas”网络也属于软竞争(soft-max)的自适应聚类算法。而这些算法的区别在于邻域核函数反映的邻域特征。SOFM邻域核函数反映了输出空间固定有序网格点间距离的特性，其缺点是无法在输入空间中刻画获胜神经元与其邻域神经元的距离信息，原因是在输出空间上的获胜神经元并不一定在原数据空间中存在相应的原像。而“Neural-Gas”网络的邻域核函数刻画了输入数据空间上某数据点 $\boldsymbol{x}_t, t \in \{1,2,\cdots,M\}$ 与特征空间上所有聚类权矢量 $\boldsymbol{\omega} = \{\boldsymbol{w}_1, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_N\}$ 间距离的排列顺序，所以聚类结果能够在输入空间中获得直观的解释，能够更好地保持特征空间的拓扑有序性。“Neural-Gas”网络的代价函数为

$$E(\boldsymbol{\omega}) = \frac{1}{2C(\lambda)} \sum_{t=1}^{M} \sum_{j=1}^{N} h_\lambda(k_j(\boldsymbol{x}_t, \boldsymbol{\omega})) \left\| \boldsymbol{x}_t - \boldsymbol{w}_j \right\|^2 \tag{1}$$

其中 $h_\lambda(k_j(\boldsymbol{x}_t, \boldsymbol{\omega})) = \exp(-k_j(\boldsymbol{x}_t, \boldsymbol{\omega})/\lambda)$，$k_j(\boldsymbol{x}_t, \boldsymbol{\omega})$ 是 $\boldsymbol{w}_l$ 满足 $\left\| \boldsymbol{x}_t - \boldsymbol{w}_l \right\| < \left\| \boldsymbol{x}_t - \boldsymbol{w}_j \right\|$ 的个数，即邻域距离的排列序号，若距离越近，排列序号越小，$h_\lambda(k_j(\boldsymbol{x}_t, \omega))$ 就越大；λ 从较大常数 λ_0 开始，随迭代次数的增加单调递减，即 $\lambda = \lambda_0 (\lambda_1 / \lambda_0)^{t/t_{\max}}$，其目的是通过构造该可微代价函数在随 λ 值的减小时逐渐地一致逼近硬 K 均值不可微函数以保证该算法尽可能接近(甚至达到)全局最优，其作用与确定性退火(DA)算法温度参数相同；归一化因子 $C(\lambda) = \sum_{j=1}^{N} h_\lambda(k_j(\boldsymbol{x}_t, \omega))$，仅与 λ 有关。令 $P_\lambda(k_j(\boldsymbol{x}_t, \omega)) = h_\lambda(k_j(\boldsymbol{x}_t, \omega))/C(\lambda)$，所以 $\sum_{j=1}^{N} P_\lambda(k_j(\boldsymbol{x}_t, \omega)) = 1$。结合上面的分析，可将 $P_\lambda(k_j(\boldsymbol{x}_t, \omega))$ 视为模糊隶属度函数，它说明数据空间中的任一数据向量 $\boldsymbol{x}_t$ 是以一定概率比值模糊分配给所有权向量，该分配概率与数据向量 $\boldsymbol{x}_t$ 到所有权向量的距离的排列顺序有关，因此“Neural-Gas”网络可视为一种模糊聚类算法。

由式(1)导出的学习规则保证了学习算法具有收敛速度快、代价误差小，收敛性稳定等优点，但是对于非均匀分布的数据集，由于“Neural-Gas”网络的特征空间(聚类权空间)与原空间的拓扑保持，其大部分神经元权矢量非均匀地分散在原数据空间点附近，当这些神经元映射在二维输出网格上时，神经元的位置分布不能真实地反映原空间数据点的距离信息，所以“Neural-Gas”网络用于数据降维或数据可视化展现出较差的结构特性。

2.2 Sammon 映射[5, 6]

传统 MDS 解决了数据的降维和可视化问题，它能够保证原空间与输出空间数据点间的距离信息的相似性。MDS 刻画原数据空间和输出空间数据点间的距离误差信息(包括距离间绝对误差和相对误差)，其中 Sammon 非线性映射强调原输入空间和输出空间上数据点间距离的绝对误差和相对误差的折中。尽管 Sammon 保持了原数据空间与输出空间数据点间的距离相似特性，但其缺点包括迭代计算复杂度较大，对映射初值较敏感，易陷入局部多极值以及输入空间上数据点间距离无论长短都对代价函数的贡献相等。

3　混合 Neural-Gas 网络和 Sammon 映射的数据可视化法

签于“Neural-Gas”网络输出空间上神经元间不能保持输入空间数据点间的距离一致性以及Sammon计算复杂度较大，易陷入局部多极值等问题，提出在“Neural-Gas”网络算法的基础上引入Sammon映射惩罚项，即通过“Neural-Gas”网络预先聚类，获取原数据空间的特征权矢量，然后再通过Sammon非线性映射使特征空间上神经元的权矢量能均匀地分布在原数据空间上，其结果既保持了两空间的拓扑关系又保证了两空间的距离相似性。因“Neural-Gas”网络被视为一种模糊聚类算法，算法中任一数据向量均以一定概率比值模糊分配给所有权向量，所以由“Neural-Gas”网络和Sammon映射的代价函数来构造混合算法的代价函数时必须考虑分配概率对代价函数的影响。由此混合代价函数可表示为

$$E_t(\boldsymbol{\omega}) = \frac{1}{2} \left\| \sum_{j=1}^{N} P_\lambda(k_j(\boldsymbol{x}_t, \boldsymbol{\omega}))(\boldsymbol{x}_t - \boldsymbol{w}_j) \right\|^2 + \frac{\gamma}{8} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} P_\lambda(k_i(\boldsymbol{x}_t, \boldsymbol{\omega})) P_\lambda(k_j(\boldsymbol{x}_t, \boldsymbol{\omega})) \frac{(d_{ij}^2 - \beta \Delta_{ij}^2)^2}{\beta \Delta_{ij}^2} \tag{2}$$

其中 $\boldsymbol{\omega} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_N\}$ 是特征空间神经元权矢量集合，其神经元权间的欧氏距离 $d_{ij} = \left\| \boldsymbol{w}_i - \boldsymbol{w}_j \right\|$；2D 输出空间上神经元间的欧氏距离 $\Delta_{ij} = \left\| \mathrm{pos}_i - \mathrm{pos}_j \right\|$，$\mathrm{pos}_i$ 是输出网格神经元 i 的坐标；β 是分辨率参数，与神经元的数目和数据集的特性密切相关，其目的是为了控制特征空间与原空间数据的相对位置关系；若 β 值较大，则分辨率较低，特征空间能够完全覆盖原空间数据结构，但会导致个别神经元被浪费，若 β 值较小，则分辨率较高，神经元能被充分利用，但特征空间不能完全覆盖原空间，导致部分原始数据特性被忽略，所以为确保特征空间恰好完全覆盖整个原始数据空间而选择合适的 β 值是非常重要的[7]；γ 是两代价函数之间的调整参数，它根据数据集的特性以及混合代价函数前后两项对整体代价贡献的程度进行调整，一般取值为 1 到 3 之间。E_t 关于 $\boldsymbol{w}_i$ 的梯度为

$$\frac{\partial E_t}{\partial \boldsymbol{w}_i} = -P_\lambda(k_i(\boldsymbol{x}_t, \boldsymbol{\omega})) \sum_{j=1}^{N} P_\lambda(k_j(\boldsymbol{x}_t, \boldsymbol{\omega})) \cdot \left[(\boldsymbol{x}_t - \boldsymbol{w}_j) + \gamma(\boldsymbol{w}_j - \boldsymbol{w}_i) \frac{(d_{ij}^2 - \beta \Delta_{ij}^2)}{\beta \Delta_{ij}^2} \right] + R_i + M_i \tag{3}$$

其中 $R_i = \frac{1}{C(\lambda)^2} \sum_{j=1}^{N} \frac{\partial h_\lambda(k_j(\boldsymbol{x}_t, \boldsymbol{\omega}))}{\partial \boldsymbol{w}_i} \sum_{k=1}^{N} h_\lambda(k_k(\boldsymbol{x}_t, \boldsymbol{\omega}))(\boldsymbol{x}_t - \boldsymbol{w}_k)^{\mathrm{T}} \cdot (\boldsymbol{x}_t - \boldsymbol{w}_j)$；$M_i = \frac{\gamma}{4C(\lambda)^2} \sum_{j=1}^{N} \frac{\partial h_\lambda(k_j(\boldsymbol{x}_t, \boldsymbol{\omega}))}{\partial \boldsymbol{w}_i} \sum_{k=1, k\neq j}^{N} h_\lambda(k_k(\boldsymbol{x}_t, \boldsymbol{\omega}))$

$\cdot\dfrac{(d_{jk}^2-\beta\Delta_{jk}^2)^2}{\beta\Delta_{jk}^2}$。

令 $l_j=\dfrac{1}{C(\lambda)^2}\sum_{k=1}^{N}h_\lambda(k_k(\boldsymbol{x}_t,\boldsymbol{\omega}))(\boldsymbol{x}_t-\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{x}_t-\boldsymbol{w}_j)$，$d_j^2=\left\|(\boldsymbol{x}_t-\boldsymbol{w}_j)\right\|^2, j\in\{1,2,\cdots,N\}$，则 $R_i=\sum_{j=1}^{N}\dfrac{\partial h_\lambda(k_j(\boldsymbol{x}_t,\boldsymbol{\omega}))}{\partial\boldsymbol{w}_i}l_j=\sum_{j=1}^{N}\dfrac{\partial h_\lambda(k_j(\boldsymbol{x}_t,\boldsymbol{\omega}))}{\partial k_j(\boldsymbol{x}_t,\boldsymbol{\omega})}\dfrac{\partial k_j(\boldsymbol{x}_t,\boldsymbol{\omega})}{\partial\boldsymbol{w}_i}l_j$。

由于 $k_j(\boldsymbol{x}_t,\omega)$ 是 $\boldsymbol{w}_l$ 满足 $\left\|\boldsymbol{x}_t-\boldsymbol{w}_l\right\|^2<\left\|\boldsymbol{x}_t-\boldsymbol{w}_j\right\|^2$ 的个数，即满足 $d_l^2<d_j^2$ 的个数。定义函数 $u(d_j^2-d_l^2)=\begin{cases}1, & d_j^2>d_l^2\\ 0, & d_j^2\le d_l^2\end{cases}$，所以 $k_j(\boldsymbol{x}_t,\boldsymbol{\omega})=\sum_{l=1}^{N}u(d_j^2-d_l^2)$。在求 $\dfrac{\partial k_j(\boldsymbol{x}_t,\boldsymbol{\omega})}{\partial\boldsymbol{w}_i}$ 时，要求 $k_j(\boldsymbol{x}_t,\boldsymbol{\omega})$ 项必须包括 d_i^2 项，即满足条件 $d_j^2\ge d_i^2$。令 $u(d_j^2-d_l^2)$ 的导数为 $\delta(d_j^2-d_i^2)$，所以 R_i 可表示为

$$R_i=\frac{\partial h_\lambda(k_i(\boldsymbol{x}_t,\boldsymbol{\omega}))}{\partial k_i(\boldsymbol{x}_t,\boldsymbol{\omega})}\,l_i\sum_{l=1}^{N}\delta(d_i^2-d_l^2)-\sum_{j=1}^{N}\frac{\partial h_\lambda(k_j(\boldsymbol{x}_t,\boldsymbol{\omega}))}{\partial k_j(\boldsymbol{x}_t,\boldsymbol{\omega})}\,l_j\,\delta(d_j^2-d_i^2)$$

因 $\delta(d_j^2-d_i^2)\begin{cases}\neq 0, & d_j^2=d_i^2\\ =0, & d_j^2\neq d_i^2\end{cases}$，且 $\sum_{j=1}^{N}\delta(d_j^2-d_i^2)=\sum_{l=1}^{N}\delta(d_i^2-d_l^2)$，所以可得 $R_i=0$。同理可证 $M_i=0$。基于上述的分析，特征空间上各神经元权值的学习规则为

$$\Delta\boldsymbol{w}_i=\varepsilon(t)\cdot P_\lambda(k_i(\boldsymbol{x}_t,\boldsymbol{\omega}))\cdot\left\{\sum_{j=1}^{N}P_\lambda(k_j(\boldsymbol{x}_t,\boldsymbol{\omega}))\left[(\boldsymbol{x}_t-\boldsymbol{w}_j)+\gamma(\boldsymbol{w}_j-\boldsymbol{w}_i)\frac{(d_{ij}^2-\beta\Delta_{ij}^2)}{\beta\Delta_{ij}^2}\right]\right\}\quad(4)$$

为保证特征空间上神经元权值的更新从混合算法到“Neural-Gas”网络的平滑过渡，引入单调递减函数 $\xi_t\in(0,1)$，并令 $h_i=P_\lambda(k_i(\boldsymbol{x}_t,\boldsymbol{\omega}))$，将式(4)修正为

$$\boldsymbol{w}_i(t+1)=\boldsymbol{w}_i(t)+\varepsilon(t)\cdot h_i\cdot\sum_{j=1}^{N}h_j\left\{(\boldsymbol{x}_t-\boldsymbol{w}_j)+\gamma(\boldsymbol{w}_j-\boldsymbol{w}_i)\left[\xi_t+(1-\xi_t)\frac{(d_{ij}^2-\beta\Delta_{ij}^2)}{\beta\Delta_{ij}^2}\right]\right\}\quad(5)$$

该混合算法是在“Neural-Gas”算法的基础上引入修正的Sammon惩罚项来保证原数据空间与输出空间数据点间的距离相似性。当两空间数据点间的距离调整到相似后，混合算法平滑地恢复为原“Neural-Gas”算法。因原“Neural-Gas”算法能够稳定地收敛到全局最优或接近全局最优[1]，结合上面的分析，在适宜的参数下混合算法也能稳定地收敛到全局最优或接近全局最优。另外，与传统 Sammon 算法相比，该混合算法仅计算特征神经元权矢量的距离信息，所以在较大程度上减小了计算复杂度，提高了映射的收敛速度。

4 数据仿真

为了比较和验证“Neural-Gas”网络、Sammon映射和混合算法间的性能，分别用3D线性可分数据样本集和Iris线性不可分数据样本集进行测试，3D线性可分数据样本集由均值分别为 $[5.0,7.0,6.0]^{\mathrm{T}}$，$[-2,5.0,-3.0]^{\mathrm{T}}$，$[-10,6.0,2.0]^{\mathrm{T}}$ 满足均匀分布，数据点数为50的3组数据组成。另外，为了定量比较这些算法间性能，引入相对标准方差(RSD)[7]，表征将输入空间上任一数据点到其邻域内各数据点的距离分别与输出空间上对应的距离相比所得比值分布的偏离程度以刻画两空间上数据间距离的相似程度。若两空间上对应的任一数据点与其邻域内各数据点的距离比值均相等，则方差RSD等于0，性能最好；若部分距离比不相等会导致RSD较大，性能较差，即RSD越小，性能越好。设神经元数N分别取400或100，$\lambda=N/2\sim0.125$，$\varepsilon=0.99\sim0.01$，而 β 和 γ 值视不同数据集的分布与神经元权值的关系进行适当调整。图1–图6的结果表明，混合算法和Sammon映射对这两类数据样本集的可视化均能真实地刻画原空间数据点的距离信息，而“Neural-Gas”网络却不能，证实了混合算法的有效性。另外，根据表1的结果，在不同的初始权值和各类数据样本集的情况下混合算法的 RSD 值均小于另两种算法，表明混合算法的可视化具有紧致性和稳定性。尽管Sammon映射的可视化能真实地刻画原空间数据点的距离信息，但它的RSD值对初始权值的设定较为敏感，所以Sammon映射的性能不稳定。综上所述，混合算法是可行，有效而且较稳定。

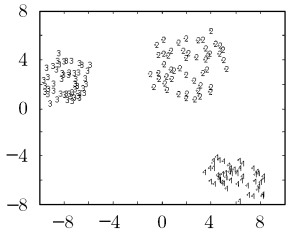

图 1 3D 数据集的 Sammon 的可视化图

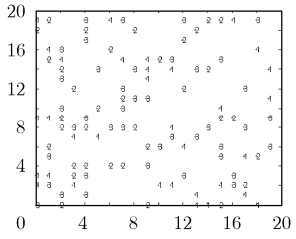

图 2 3D 数据集的 Neural-Gas 的可视化图

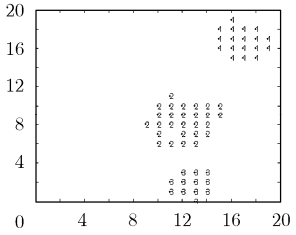

图3 3D数据集的混合算法的可视化图

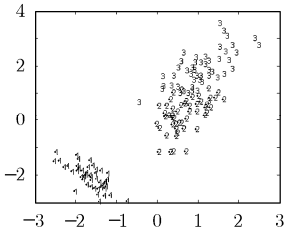

图4 Iris数据集的Sammon可视化图

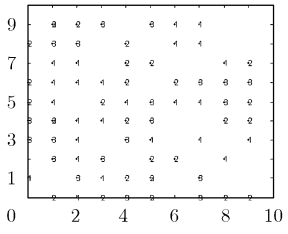

图 5 Iris 数据集的 Neural-Gas 可视化图

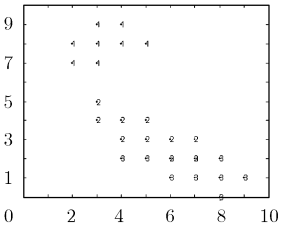

图 6 Iris 数据集的混合算法可视化图

表 1　3 种算法的 RSD 比较

数据样本集	混合算法	Sammon 算法	Neural-Gas 网络
3D 合成数据集	0.07	0.12	1.50
Iris 数据集	0.03	0.09	1.25

5　结束语

针对传统的“Neural-Gas”网络在对数据进行可视化时在输出空间上表现出极不理想的距离相似特性，提出一种“Neural-Gas”网络与Sammon映射相结合的可视化算法。该算法既降低了计算复杂度，又保持了输入空间与输出空间上数据点间的距离相似性。仿真结果在适宜的参数值(如β和γ等)下验证了混合算法的可行性和有效性。结果表明，混合算法对合成数据集或现实数据集的可视化性能均优于“Neural-Gas”网络和Sammon映射。因参数的选择直接影响到最终的可视化性能，所以如何根据混合算法的收敛性和稳定性选择较佳参数的问题还有待从理论上深入讨论。

参 考 文 献

[1] Martinetz T M and Berkovioch S G . Neural-Gas network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Network*, 1993, 4(4): 558–568.

[2] Claussen J C and Villmann T. Magnification control in winner relaxing neural gas. *Neurocomputing*, 2004, 63(2): 125–137.

[3] Kong A. Interactive visualization and analysis of hierarchial neural projections for data mining. *IEEE Trans. on Neural Network*, 2000, 11(3): 615–624.

[4] Pal N R and Eluri V K. Two efficient connectionial schemes for structure preserving dimensionality reduction. *IEEE Trans. on Neural Network*, 1998, 9(6): 1142–1154.

[5] Deodhare D and Kesheorey A. An improved sammon's nonlinear mapping algorithm. Proceedings of the International Conference on cognition and recongition, Karnataka city, India, Dec. 2005: 74–82.

[6] Sammon J W. A nonlinear mapping for data structure analysis. *IEEE Trans. on Comput.*, 1969, 18(5): 401–409.

[7] Wu Sitao and Chow W S. PRSOM: A new visualization method by hybridizing multi-dimensional scaling and self-organizing map. *IEEE Trans. on Neural Networks*, 2005, 16(6): 1362–1380.

晋良念：　男，1974年生，博士生，研究方向为数据挖掘、自适应信号处理、神经网络.

欧阳缮：　男，1960年生，教授，博士生导师，研究方向为自适应信号处理、通信信号处理及神经网络等.