计算机科学
COMPUTER SCIENCE

基于改进SOM网络的聚类算法

蒋锐，范姝文，王小明，徐友云

引用本文

蒋锐，范姝文，王小明，徐友云. 基于改进SOM网络的聚类算法[J]. 计算机科学，2025，52(8)：162-170.

JIANG Rui, FAN Shuwen, WANG Xiaoming, XU Youyun. Clustering Algorithm Based on Improved SOM Model [J]. Computer Science, 2025, 52(8): 162-170.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700017

基于改进 SOM 网络的聚类算法

蒋 锐 范姝文 王小明 徐友云
南京邮电大学通信与信息工程学院 南京 210003

摘 要 在自组织映射(Self-organizing Map,SOM)模型的训练过程中,不同类数据对权重矩阵的更新有不同作用,某一类数据对权重矩阵的更新会对其他类获胜神经元特征向量产生偏离其数据特征的影响,从而降低算法聚类精度。针对以上问题,提出一种改进的基于置信度 SOM 模型(Improved Confidence-based SOM Model,icSOM)。样本数据首先由 K-means 算法初步分类,为模型训练提供更多的数据信息;然后将预分类后的数据分别训练相互独立的 SOM 模型,以消除不同类之间的影响;最后在传统 SOM 模型基础上提出置信度矩阵概念,通过综合判断获胜神经元的置信度及其与输入数据间的欧氏距离最终得到置信神经元,根据置信神经元所属类别给数据分配聚类标签。在鸢尾花数据集(Iris)及葡萄酒数据集(Wine)上利用 icSOM 进行聚类分析,实验结果表明,所提算法可以更好地处理样本数据,取得了较好的聚类效果。

关键词:机器学习;无监督学习;聚类;自组织特征映射神经网络
中图分类号 TP181

Clustering Algorithm Based on Improved SOM Model

JIANG Rui,FAN Shuwen,WANG Xiaoming and XU Youyun
School of Communications and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China

Abstract In the training process of the Self-Organizing Map,different classes of data have varying effects on the update of the weight matrix. Therefore,the update of the weight matrix for a certain class of data will have an impact on the feature vectors of the winning neurons,which are corresponding to other classes of data. This impact causes the winning neurons to deviate from the features of the data,thus reducing the clustering accuracy of the algorithm. Regarding the above issue,this paper proposes an improved confidence-based SOM model(icSOM). Firstly,the sample data is classified by the K-means algorithm to provide more information for model training. Secondly,the pre-classified data is used for training different classes SOM models to eliminate the influence caused by data from different classes. Based on the traditional SOM model,the concept of confidence matrix is then proposed. By comprehensively evaluating the confidence of the winning neurons and their Euclidean distance to the input data,the confident neuron is finally obtained. The clustering label that assigned to this input data is same as this confident neuron's class. Using icSOM for clustering analysis of the Iris dataset and the Wine dataset,the experimental results show that the proposed algorithm can handle sample data more effectively and achieve better clustering performance.

Keywords Machine learning,Unsupervised learning,Clustering,Self-organizing feature map neural network

1 引言

随着大数据时代的到来,作为人工智能(AI)[1-2]的子集之一,机器学习(Machine Learning,ML)受到广泛关注。ML 这一概念由 Samuel 提出,其核心思想是无需对计算机进行精确编程就可使其具有学习能力[3-4]。ML 模拟人类学习行为,使计算机获取新的知识和技能,重新组织已有的模型结构并优化自身性能[5],是一种可以基于过去预测未来的方法,即将历史数据及先验知识作为训练数据来训练模型,然后使用模型对测试数据进行预测并不断优化算法自身[6]。模型是 ML 的基本要素,构建具有准确预测数据能力的模型,离不开大量具有代表性的训练数据[7]。根据训练数据是否带有标签信息,可以将 ML 分为有监督学习(Supervised Learning)[8]和无监督学习(Unsupervised Learning)[9]两大类。

有监督学习是指利用具有标签的数据训练模型,即已知输入变量和输出变量,通过建立输入输出的对应关系来预测测试数据的输出变量。若输出变量是对真实值的逼近(即定量分析),则为回归问题;若输出变量是一种类别(即定性分析),则为分类问题。有监督学习中的经典算法[10]包括:K 近邻算法(K-Nearest Neighbor,KNN)[11]、决策树(Decision Trees,DTs)、朴素贝叶斯(Naive Bayes,NB)等。

无监督学习是指利用只有特征向量且不携带标签信息的样本数据训练模型,其最大特点是从数据出发理解数据本身,分析数据的目的不局限于分类,而是按照不同维度的特征寻

到稿日期:2024-07-05 返修日期:2024-10-26
基金项目:国家自然科学基金(62371246)
This work was supported by the National Natural Science Foundation of China(62371246).
通信作者:蒋锐(j_ray@njupt.edu.cn)

找数据规律，这与有监督学习中通过已有标签的数据集去训练得到一个最优模型是不同的，同时，这也是无监督学习应用更为广泛的原因。无监督学习常用于对数据进行降维和聚类。降维[12]是压缩数据信息的过程，即在尽量保留数据结构的同时去除冗余数据以降低数据复杂度，主要算法有主成分分析法（Principal Component Analysis，PCA）[13]、奇异值分解法等。聚类是将未被标记的数据进行自动分类，常见聚类算法有 K 均值聚类（K-means）[14]、DBSCAN 聚类[15]、层次聚类等。

自组织映射（SOM）[16-19]是无监督学习中的经典聚类算法，近年来被广泛应用于各个领域，如数据可视化、聚类分析、对高维数据进行特征提取、异常检测、为其他无监督学习提供数据预处理等。国内外学者不断对其进行深入研究以提升算法性能。Zhou 等[20]提出了一种基于 SOM 网络的遥感影像分类算法，该算法将像素转换为“量子像素”，并提出用 PE 系数表征量子像素间的关系，建立 SOM 模型对量子像素间的关联性进行分析，以实现图片分类。该算法有效改善了传统 SOM 模型只考虑神经元之间的欧氏距离而不能真实反映遥感影像中像素的类关系的问题，但该算法的聚类效果十分依赖初始数据，不同的遥感影像数据使得模型的最优参数难以确定，影响了分类效果。Li 等[21]提出了一种自监督的自组织聚类网络（Self-supervised Self-organizing Clustering Network，OCNet），该网络将特征提取和聚类过程相结合，将自组织层的权重作为聚类中心，将自组织层的输出作为特征与聚类中心之间的相似度，级联多层自组织层以实现在不同特征维度下的多个聚类空间的特征聚类。该算法解决了传统聚类中特征提取和聚类过程相互独立导致的特征提取时没有考虑如何促进聚类的问题，但该算法必须设定超参数以限制聚类簇的大小，否则会产生大型聚类，且该算法要求不同类别的数据分布尽量均衡，因此，在处理特征分布不均匀的数据集时聚类效果不理想。Yan 等[22]提出了一种改进自组织映射算法，该算法先用小波变换去除数据中的噪声以降低对聚类效果的干扰，然后对去噪数据进行特征提取并将其作为初始权重矩阵用于后续 SOM 模型的聚类分析，此外，通过设置输出层各神经元之间的映射权重，增强了输入数据与各节点权重之间的关系，减少了未充分利用或完全未使用的神经元。该算法可用于处理规模较大且含有噪声的幂时序数据集，且一定程度地避免了传统 SOM 模型中权重分配不均匀持续累加导致某些神经元始终不能获胜的问题，但其修改了神经元权重分布规则，使得高权重神经元不易获得更高权重，因此在处理具有明显特征差异的数据时，算法的收敛速度慢且时间复杂度高。Xie 等[23]提出了 SOM-AdaDBSCAN 算法用于对数据进行二次聚类，初始数据首先由 SOM 模型初步聚类，然后根据每个类集群的大小对事件进行比例抽样以减少训练样本的数量，最后结合 SNN 密度和欧氏距离来减少训练样本的数量，以形成改进的 DBSCAN 算法 AdaDBSCAN 用于辅助聚类。该算法解决了传统 DBSCAN 难以处理多密度聚类、第一步聚类计算速度慢的问题，但用于完成一次聚类的 SOM 模型增加了算法复杂度，且初步聚类后按类群大小对样本进行抽样，损失了部分信息量，造成了误差传递。Khan 等[24]提出的混合 SOM 算法将传统 SOM 模型与 KNN 模型相结合，在含有少量噪声的健康数据集上训练 SOM 算法，拟合健康数据后，剔除过于稀疏或低于阈值的节点，避免获胜神经元被噪声污染。该算法在 SOM 模型之后增加一层 KNN 模型，基于质心与观测数据点之间的欧氏距离对数据进行分类。其具有良好的可扩展性，通过阈值判断减少算法输入输出数量及运算时间，降低了算法复杂度，但剔除节点可能会损失部分有用的数据信息。Bendjama 等[25]提出 PCA-SOM 算法，该算法先利用 PCA 将数据矩阵投影到较小的子空间中，以降低原始数据维度；然后，通过建立数据间的统计相关性，将更能代表数据特征的特性作为 SOM 模型的输入；最后，SOM 根据数据间的欧氏距离将其分类。相较于直接减少输入数据数量的做法，该联合算法提高了不同类别数据间的区分度以及算法的准确性，但在处理低维数据时，维度的减少在一定程度上会影响数据处理的准确度。

传统 SOM 模型属于无监督学习，具有很强的自适应性，通过输出层中各神经元的竞争可以实现将高维数据特征映射到低维空间，保持拓扑结构，但其也有一定的局限性，如当训练样本的类别较少时，聚类结果会受到不同类别输入的先后顺序的影响，在没有完成学习之前，不能加入新的训练数据类别，且 SOM 模型在学习训练数据用于更新权重矩阵时，不同类别的数据会相互影响。综上所述，本文提出了一种改进的基于置信度 SOM 模型（icSOM），样本数据先由 K-means 算法初步分类，然后用分类后的数据分别训练 SOM 模型，以消除不同类之间的影响。该算法可以更好地处理样本数据，通过对数据进行预分类来为传统无监督学习提供更多的数据信息量，实现了更高的聚类结果准确度。

2　SOM 模型

2.1　SOM 模型原理简介

SOM 模型通过无监督学习，将高维输入数据映射到低维空间节点上，每个神经元节点都有相应的权重向量，该权重向量表征了神经元节点在输入层的位置，特征相似的样本数据会映射到相邻的神经元节点上，使得输出层保持了输入数据的拓扑结构，反映了输入数据间的关系。因此 SOM 模型被广泛用于数据可视化、降维、聚类分析等。SOM 模型采用的 Kohonen 算法[26]，由 Kohonen 提出，其核心思想是“自组织”和“竞争”：输出层的各神经元进行竞争，最终获胜的神经元获得对输入数据的响应机会；且获胜神经元对其余神经元节点的影响符合“墨西哥草帽”函数（见图 1），获胜神经元周围的神经元节点为兴奋状态，距离获胜神经元较远的神经元节点为抑制状态，这使得与获胜神经元有关的权重向量可以朝着更有利于其竞争的方向更新。通过不断学习输入数据，输出层最终形成可以表征输入数据分布的拓扑结构。

传统 SOM 模型包含输入层与输出层，结构如图 2 所示。输入层接收样本数据，其维度与输入数据的维数相同；输出层由神经元节点构成，通过竞争得到响应输入数据的机会，也被称为竞争层。输入层与输出层的各节点之间通过权重向量全连接，权重向量表征连接强度，反映输入数据的拓扑结构。训练开始前，首先初始化权重矩阵，为每个权重向量分配随机数值；然后输入样本数据，计算其与各神经元节点间的欧氏距离，距离最小的节点为激活点，即获胜神经元；最后更新激活点及其邻域内节点的参数，以实现各权重向量的更新。每次

迭代过程中均对邻域函数和学习率进行更新，经多次迭代训练后将其输入测试数据，映射在输出层上的获胜神经元代表了该数据的聚类类别。

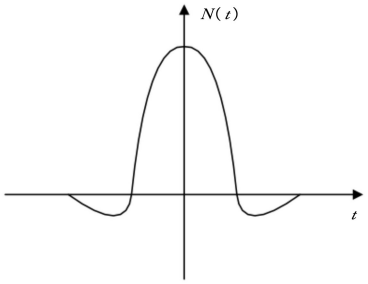


图 1 “墨西哥草帽”函数示意图

Fig. 1 Schematic diagram of “Sombrero” function

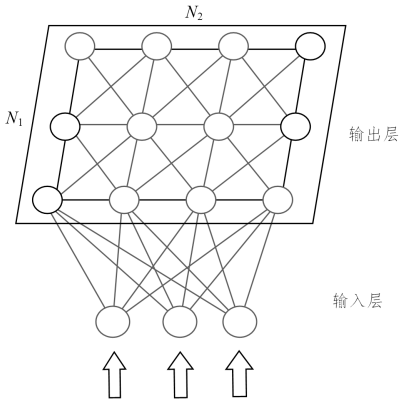


图 2 SOM 结构

Fig. 2 Architecture of SOM

2.2 SOM 模型训练过程

步骤 1 模型初始化。输入层由 D 个节点组成，其中 D 与输入特征维度相同；输出层包含 M 个神经元节点，形成二维矩阵，其中 $M=N_1 \cdot N_2$，依据经验公式，令 $N_1=N_2=\sqrt{5\sqrt{N}}$，其中 N 为训练样本数量。随机初始化各神经元节点对应的权重向量值 $w_{ij}(i=1,2,\cdots,D;j=1,2,\cdots,M)$，生成初始化权重矩阵 $\boldsymbol{W}$，其大小为 $M \cdot D$。设置初始学习率 η_0、邻域衰减率 δ_0、迭代次数 T。

步骤 2 样本数据预处理。输入训练数据并进行归一化和正则化预处理，使样本数据特征有相同的度量尺度并避免数据过拟合。

步骤 3 寻找获胜神经元。选取一个样本数据 x，计算与其各神经元节点间的欧氏距离。

$$dis=\| x-w_{ij} \| \tag{1}$$

选取距离最近的点 (i_x,j_x) 作为激活点，即样本 x 映射在输出层的获胜神经元。

步骤 4 计算权重系数，更新邻域函数。令激活点的权重系数为 1，其他神经元节点通过计算与激活点间的距离得到自身的权重系数，进而得到本次迭代对应的邻域函数。

$$g(i,j)=\mathrm{e}^{-\frac{(i_x-i)^2}{2\delta^2}}\mathrm{e}^{-\frac{(j_x-j)^2}{2\delta^2}} \tag{2}$$

其中，δ 随迭代次数的增加而减小，以降低影响程度，其更新方式如下：

$$\delta=\delta_0\left(\frac{1-t}{t_{\max}}\right) \tag{3}$$

步骤 5 更新学习率。

$$\eta=\eta_0\left(\frac{1-t}{t_{\max}}\right) \tag{4}$$

步骤 6 更新权重矩阵。该过程类似于寻找聚类中心，每次更新都使 w_{ij} 更接近 x。

$$\boldsymbol{W}(t+1)=\boldsymbol{W}(t)+\eta g(x-\boldsymbol{W}(t)) \tag{5}$$

步骤 7 当满足迭代终止条件后，完成训练。此时输出层上的每个神经元的权重向量都与映射在其上的输入数据的均值无限相似，因此各神经元可以用于表征输入数据的特征。

$$\lim_{t\to T} w_{ij}(t)\approx x_{\mathrm{mean}} \tag{6}$$

SOM 模型训练流程如图 3 所示。

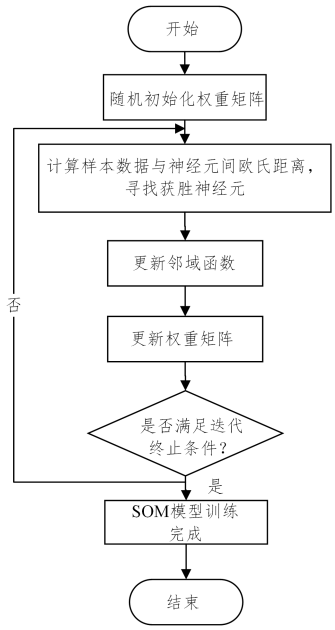


图 3 SOM 模型训练流程

Fig. 3 Flowchart of SOM model training

假设训练数据分为 3 类，在 SOM 模型训练过程中，各神经元节点的权重向量变化趋势如图 4 所示。

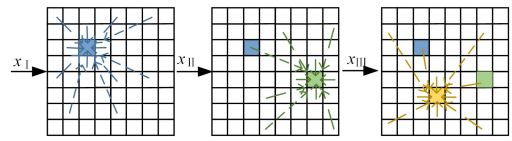


图 4 SOM 模型中神经元权重向量变化的示意图

Fig. 4 Schematic diagram of neuron weight vector changes in SOM model

图 4 中，x_{I}，x_{II}，x_{III} 分别代表不同类别的训练数据；不同的色块代表不同类别的输入数据映射在输出层的对应获胜神经元位置；线条粗细及虚实代表式(2)和式(5)的权重向量更新规则，即距离获胜神经元越近的神经元节点，其权重向量更新程度越大。由图 4 可知，不同类数据在训练模型权重矩阵的过程中会相互影响，使得某类别数据所对应的获胜神经元特征向量向其他类别数据的特征方向进行更新，在一定程度上降低了该神经元对此类别数据特征的代表性。因此，icSOM 模型以消除不同类数据之间的影响，实现更好的聚类效果。

3 icSOM 模型

icSOM 模型是在传统 SOM 模型的基础上进行创新，以

消除不同类数据在模型训练过程中对权重矩阵更新所产生的影响。训练数据先由 K-means 算法进行初步分类，为模型训练提供更多的数据信息；其次将预分类后的数据用于不同类别的 SOM 模型的训练，以消除不同类之间的影响；然后在传统 SOM 模型的基础上提出置信度矩阵概念，并综合判断获胜神经元的置信度及其与输入数据间的欧氏距离，得到置信神经元；最后根据置信神经元所属类别为数据分配聚类标签。

3.1　预分类

作为无监督学习中的经典算法，K-means 算法的原理简单且易于实现，算法复杂度低，收敛速度快，适用于对样本数据进行预处理。尽管预分类结果的精度较低，但通过简单预分类可以在一定程度上将特征相近的数据分为一类，为之后的无监督学习提供更多数据信息量。将样本数据作为 K-means 算法的输入，初始化聚类中心，每次迭代都将数据点分配给距离最近的聚类簇，然后重新计算簇中心点，随着迭代次数增加，数据完成聚类，簇中心点的特征权重表示该类数据的特征。K-means 算法迭代终止后，样本数据实现预分类。

3.2　置信度矩阵

传统 SOM 模型中的权重矩阵用于计算样本数据 x 与神经元节点间的欧氏距离，以找到对应的获胜神经元。而在 icSOM 算法中，将原有权重矩阵作为搜索矩阵，并提出置信度矩阵的概念，其内容为：为每个 SOM 模型初始化一个零矩阵并将其用于存储各神经元节点的置信度；对应数据 x 的获胜神经元置信度增加 1，其他神经元节点通过计算与获胜神经元间的距离得到自身的置信度增量。

$$\Delta c(i,j)=\mathrm{e}^{-\frac{(i_x-i)^2}{2\delta^2}}\ \mathrm{e}^{-\frac{(j_x-j)^2}{2\delta^2}} \tag{7}$$

其中，δ 为衰减率，随着训练迭代次数增加而减小。

遍历完所有训练数据后，可以得到该 SOM 模型的置信度矩阵，然后对其进行归一化处理，统一度量尺度，以便后续横向对比不同 SOM 模型的置信度矩阵。

3.3　icSOM 模型训练过程

步骤 1　对样本数据进行归一化和正则化预处理。

步骤 2　使用 K-means 算法对数据预分类。首先随机选取 k 个点作为初始聚类中心；然后将每一个样本数据分配给距其最近的中心点，并重新计算该类的中心；最后，当达到迭代终止条件后，即实现将数据预分类为 k 类，每类数据的特征可由其中心点权重向量表示。

步骤 3　利用 k 类数据分别训练 SOM 模型，得到 SOM_1，SOM_2，…，SOM_k。每个模型的训练流程如下。

步骤 3.1　初始化输出层大小，随机初始化权重矩阵 $\boldsymbol{W}$，设置初始学习率 η_0、初始邻域衰减率 δ_0、迭代次数 T，设置一个零矩阵作为初始置信度矩阵。

步骤 3.2　如式(1)所示，计算输入数据与各神经元间的欧氏距离，寻找获胜神经元。

步骤 3.3　按式(2)更新各神经元节点的邻域函数，并更新式(5)的权重矩阵。

步骤 3.4　按式(7)更新各神经元节点的置信度增量 Δc。

步骤 3.5　当满足迭代条件后，训练完成，得到各 SOM 模型对应的权重矩阵 $\boldsymbol{W}$、置信度矩阵 $\boldsymbol{C}$。

假设 K-means 算法将训练数据分为 3 类，在 icSOM 模型训练过程中，各神经元节点权重向量的变化及置信度如图 5 所示。其中，x_{I}，x_{II}，x_{III} 分别代表不同类别的训练数据。3 种类别的数据分别训练 SOM 模型得到各自的权重矩阵及置信度矩阵。图中不同颜色的圆形表示不同类别的数据映射在输出层的获胜神经元节点的范围；颜色深浅变化代表了当前神经元与该类数据特征均值的相似度；每个神经元节点上所标出的数字表示经过训练后该节点的归一化置信度。需要注意的是，图中获胜神经元位置以及置信度值仅为假设，且神经元节点的置信度值如果过小，则未在图中标注。对比图 4 发现，icSOM 模型将不同类数据分别用于训练，消除了不同类别的数据在权重矩阵更新过程中所产生的相互影响，使得获胜神经元更能表征该类数据的特征。

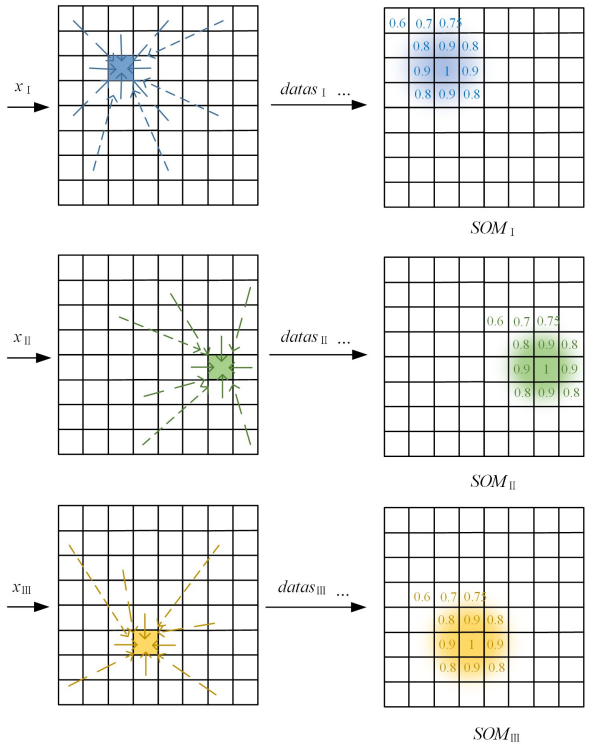


图 5　icSOM 模型中神经元权重及置信度的示意图

Fig. 5　Schematic diagram of neuron weight and confidence in icSOM model

3.4 置信神经元

同样假设数据集中有 3 类数据的情况下，首先将不同类别的数据分别用于训练，得到 SOM_{I}，SOM_{II} 和 SOM_{III}。如图 6 所示，当某一测试数据输入模型中时，红色标记代表该测试数据分别在 3 个 SOM 模型输出层上映射所得的获胜神经元 $winner_{\mathrm{I}}$，$winner_{\mathrm{II}}$ 和 $winner_{\mathrm{III}}$ 的位置，同时可知测试数据与获胜神经元之间的距离分别为 dis_{I}，dis_{II} 和 dis_{III}；映射获胜神经元所对应的置信度 c_{I}，c_{II} 和 c_{III}。此时，假设有：

$$\begin{cases} dis_{\mathrm{II}}<dis_{\mathrm{III}}<dis_{\mathrm{I}} \\ c_{\mathrm{II}}<c_{\mathrm{III}}<c_{\mathrm{I}} \end{cases} \tag{8}$$

观察发现，虽然测试数据与获胜神经元 $winner_{\mathrm{II}}$ 之间距离最短，代表该测试数据与获胜神经元 $winner_{\mathrm{II}}$ 相似度最高，但是其对应置信度 c_{II} 却最低，代表该获胜神经元 $winner_{\mathrm{II}}$ 并不能很好地表征第 II 类数据特征，因此不能将测试数据聚为此类；同理，虽然测试数据映射的获胜神经元 $winner_{\mathrm{I}}$ 对应的

置信度 c_{I} 最高，代表该映射获胜神经元 $winner_{\mathrm{I}}$ 能最好地表征第 I 类数据特征，但是测试数据与获胜神经元 $winner_{\mathrm{I}}$ 之间距离最远，代表该测试数据不符合第 I 类数据特征，因此也不能将测试数据聚为此类。最终，在综合考虑数据与获胜神经元之间的相似度及获胜神经元的置信度后，将该测试数据聚类为第Ⅲ类。

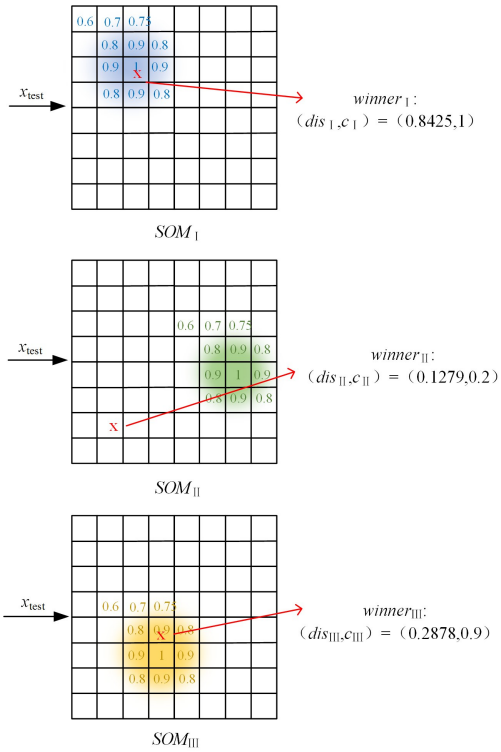


图 6 icSOM 模型中寻找置信神经元的示意图(电子版为彩图)

Fig. 6 Schematic diagram of finding confident neurons in icSOM model

综上所述，为了更好地实现测试数据聚类，综合考虑数据与获胜神经元之间的相似度及获胜神经元的置信度，在 icSOM 算法中提出置信神经元。将测试数据输入训练好的不同类别的 SOM 模型中，根据搜索矩阵得到在当前 SOM 模型中对应的获胜神经元，同时可由置信度矩阵得到该神经元对应的置信度；设定合理的置信度阈值，筛选保留符合条件的置信度所属的神经元，再比较测试数据与保留神经元间的欧氏距离，选取距离最近的节点作为该测试数据的置信神经元，将该神经元所在的 SOM 模型类别作为该数据的聚类结果。

3.5 icSOM 模型

传统 SOM 模型在训练过程中，在更新权重矩阵时不同类别数据彼此间产生影响，式(2)展示了各神经元节点通过计算与获胜神经元间的欧氏距离得到邻域函数，并将其用于式(5)中各节点权重向量的更新。而 icSOM 算法使用经过预分类的数据分别训练属于不同类别数据的 SOM 模型，消除了上述影响，同时其通过为传统无监督学习提供更多的数据信息量来改善聚类效果，缩小了无监督算法与有监督算法间的信息差。此外，传统 SOM 模型对于模糊点难以判定，当一个训练样本与多个神经元节点的欧氏距离相等时，该数据是属于不同类的边界点，传统方法只能随机选取其中一个节点作为获胜神经元，这种做法会产生判决误差。icSOM 算法基于置信度与欧氏距离综合判定输入数据所对应的置信神经元，可以解决上述问题，降低因数据模糊而引起的聚类误差。icSOM 模型训练过程及聚类分析数据过程如图 7 所示。

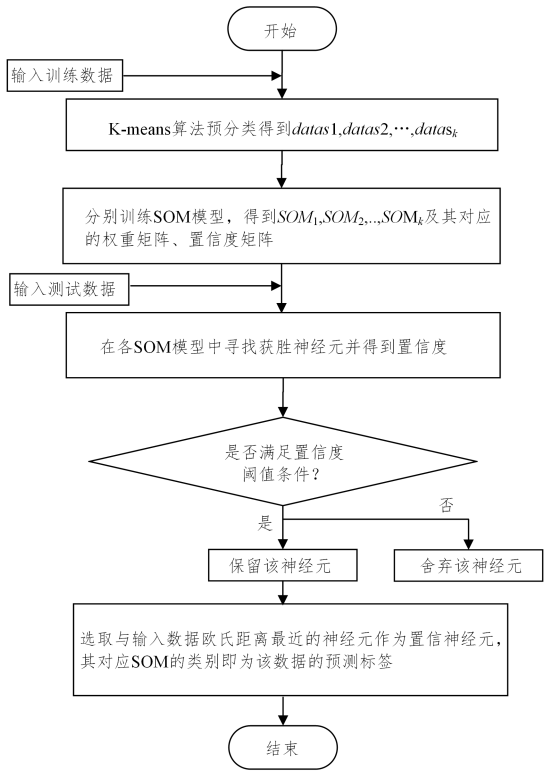


图 7 icSOM 模型流程图

Fig. 7 Flowchart of icSOM model

4 实验分析

4.1 实验设置

本文借助 PyCharm 开发工具，使用 Python 编程语言进行仿真实验。导入鸢尾花数据集(Iris)[27]及葡萄酒数据集(Wine)[28]作为待分析数据，按 7∶3 的比例将其随机划分为训练数据与测试数据。将 icSOM 算法分别与传统 SOM 算法、K-means 算法、混合 SOM 算法以及 PCA-SOM 算法进行对比，使用散点图将聚类结果可视化，可更直观地观察不同样本数据之间的聚类关系以及各算法的聚类效果。采用内部评价指标轮廓分数(Silhouette Score，S)，外部评价指标调整兰德指数(Adjusted Rand Index，ARI)、F1、准确度(Accuracy，ACC)来分析模型聚类效果及算法复杂度，并以平均运算时间(Average Operation Time，AOT)表征算法复杂度，对算法性能进行综合分析。

Iris 中包括 150 个样本且均匀分为 3 类，每个样本由 4 个特征数据组成，分别表征该样本的花萼长度、花萼宽度、花瓣长度、花瓣宽度 4 个特征；Wine 中包括 178 个样本且不均匀地分为 3 类，每个样本由 13 个特征数据组成，分别代表该样本的酒精、苹果酸、灰、灰分的碱度等成分。这两个数据集中均已给出各样本的真实标签并以 0，1，2 作为标识。首先分析 Iris 及 Wine 中各数据的特征向量，将训练数据聚类；然后通过引入数据的真实标签，为其映射在输出层上的获胜神经元分配聚类标签；最后输入测试数据，经算法处理后得到聚类标签，将其与该数据真实标签进行比较，计算各外部评价指标，

进而分析算法性能。

内部评价指标是当数据真实标签未知时，仅利用由聚类结果得到的相关统计特性来评估算法性能。轮廓分数通过计算样本与同类簇内其他数据间的平均距离 a，以及样本与邻近簇中各数据点的距离 b，来计算每个样本数据与自身聚类簇和邻近聚类簇间的紧密度和相似性关系，其取值范围为[−1,1]。轮廓分数数值越高，说明簇内紧凑且簇间数据差异越大，模型性能越好，其计算式如下：

$$s=\frac{b-a}{\max(a,b)} \tag{9}$$

$$S=\frac{1}{N}\sum_{i=1}^{N}s(i) \tag{10}$$

其中，N 为聚类簇数。

外部评价指标是指通过计算预测标签与数据真实标签的符合程度来评估算法性能。ARI 是在兰德指数的基础上进行调整的，使其可以评估簇数量大于 2 的聚类算法将数据点分配到聚类簇的准确程度，取值范围为[−1,1]，这使得聚类结果与真实情况有更高的区分度，其值越大表示聚类结果越准确，值越接近 0 表示结果越接近随机聚类。其计算式如下：

$$RI=\frac{a+b}{C_2^n} \tag{11}$$

$$ARI=\frac{(RI-E(RI))}{(\max(RI)-E(RI))} \tag{12}$$

其中，a 代表在真实标签和预测标签中都属于同一簇的样本对数；b 代表在真实标签和预测标签中不属于同一簇的样本对数；C_2^n 表示数据集中数据可以组成的总对数。

F1 指标是精确度(Precision，P)和召回率(Recall，R)的加权调和平均值，它综合考虑了 P 值对算法误判率的衡量以及 R 值对正类样本识别率的衡量。其计算式如下：

$$P=\frac{TP}{TP+FP} \tag{13}$$

$$R=\frac{TP}{TP+FN} \tag{14}$$

$$F1=\frac{2PR}{P+R} \tag{15}$$

其中，TP 为被预测为正类的正样本；FP 为被预测为正类的负样本；FN 为被预测为负类的正样本。

ACC 是指所有预测正确的样本数与全部样本数的比例，它衡量了模型的整体预测准确程度，其计算式如下：

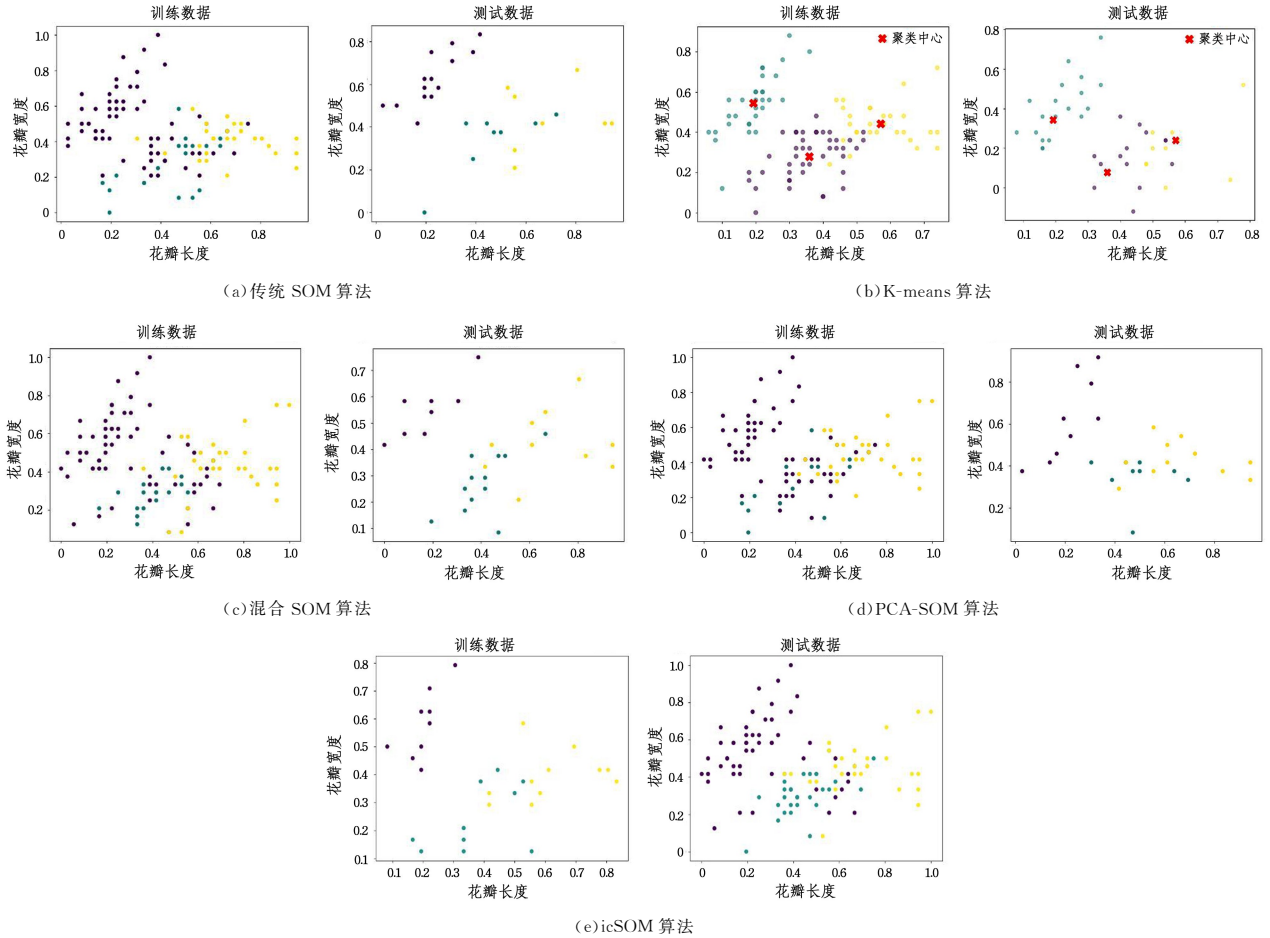
$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

其中，TN 为被预测为负类的负样本。

AOT 是指模型处理每个测试数据所用的平均时间，该指标用于衡量算法复杂度。

4.2　聚类结果可视化

图 8 分别展示了传统 SOM 算法、K-means 算法、混合 SOM 算法、PCA-SOM 算法以及 icSOM 算法在 Iris 的训练集和测试集上的聚类结果。



(a)传统 SOM 算法　(b)K-means 算法

(c)混合 SOM 算法　(d)PCA-SOM 算法

(e)icSOM 算法

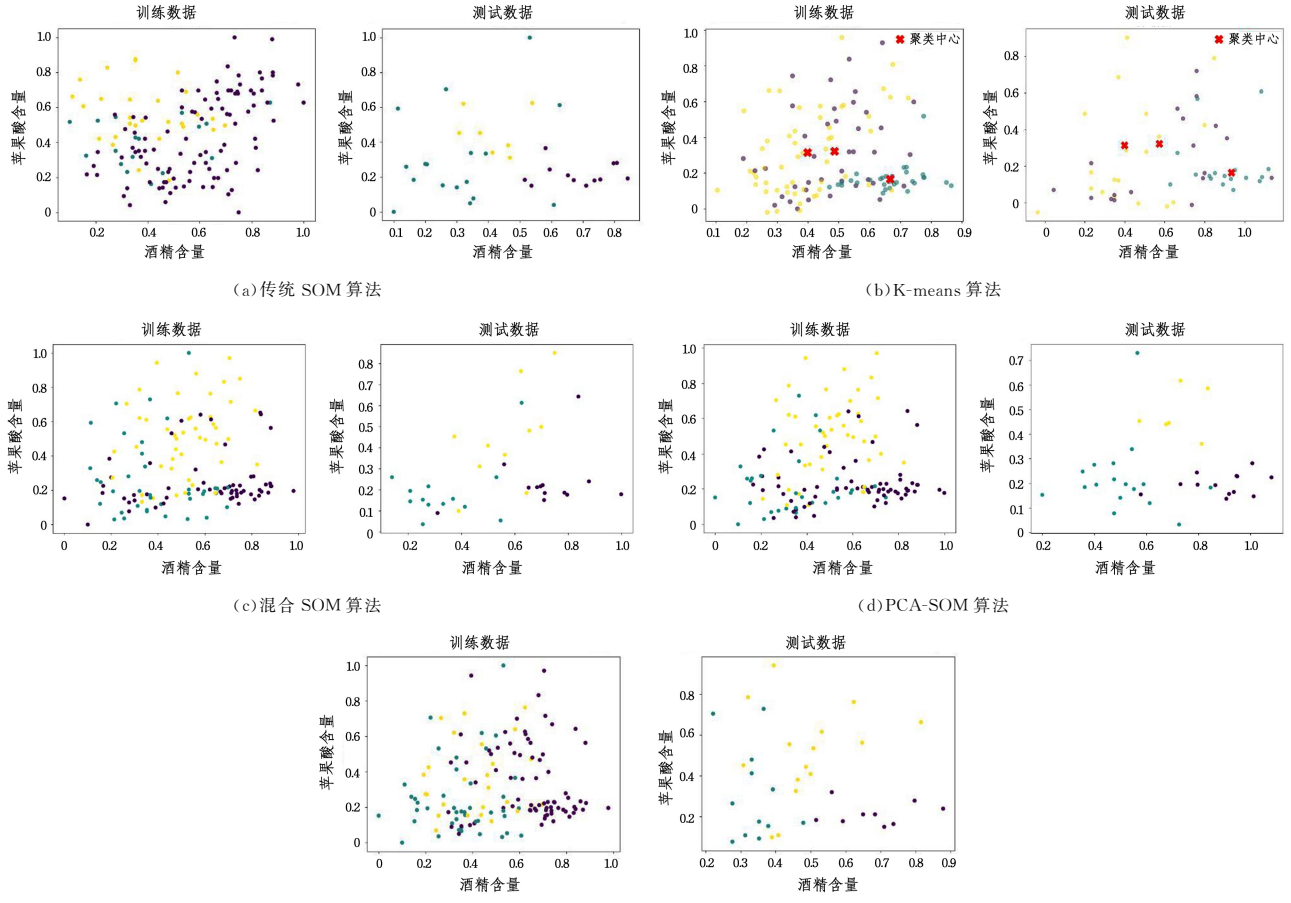
图 8　不同算法在 Iris 上的聚类结果可视化(电子版为彩图)

Fig. 8　Visualization of clustering results of different algorithms on Iris

图 8 中不同颜色的点代表不同的数据类别；图 8(b)中，红色叉号代表 K-means 算法中的聚类簇中心节点；横轴和纵轴分别代表花瓣长度和花瓣宽度，不同的特征数据可以通过横纵坐标表示在散点图的相应位置，选取这两个特性表征数据是因为通过计算 Iris 数据集中 4 个特性与类的线性相关性的皮尔逊相关系数后，发现花瓣长度和花瓣宽度这两个特性的类相关程度高，更能表征类。

图 9 分别展示了各算法在 Wine 的训练集和测试集上的聚类结果。通过计算 Wine 中 13 个特性与类的线性相关性的皮尔逊相关系数可得，葡萄酒中的酒精和苹果酸成分更能表征类，因此选择这两个特征作为散点图的横纵轴。

通过观察图 9 可知，用类相关性高的特征来表征样本数据时，可通过其在散点图中的分布位置展示不同类别数据的特性。用训练集训练模型时，不同的颜色代表了不同的种类，输入测试数据检验其聚类效果后可以发现：测试数据是由与其特征近似的训练数据的颜色所表示，即测试数据会聚为与其特性相同的训练数据的同一类，这与实验结果得到的评价指标值基本相符。散点图中呈现的点数量略少于实验代码设置的训练数据及测试数据的数量，这是由于 Iris 中不同数据的花瓣长度和花瓣宽度(或 Wine 中不同数据的酒精含量和苹果酸含量)极其接近或完全一致时，它们会在散点图上呈现于同一位置，且由于特征一致，它们会聚类为同一类别，在散点图中呈现相同的颜色。



(a)传统 SOM 算法 (b)K-means 算法

(c)混合 SOM 算法 (d)PCA-SOM 算法

(e)icSOM 算法

图 9 不同算法在 Wine 上的聚类结果可视化(电子版为彩图)

Fig. 9 Visualization of clustering results of different algorithms on Wine

4.3 聚类效果分析

表 1、表 2 列出了传统 SOM 算法、K-means 算法、混合 SOM 算法、PCA-SOM 算法以及 icSOM 算法对 Iris 和 Wine 数据集进行聚类分析后得到的各评价指标结果。传统 SOM 模型通过寻找获胜神经元并更新权重矩阵，将输入数据拓扑结构映射到输出层，特征相似的数据被聚类为同一簇。输入测试数据后，根据权重矩阵得到对应的获胜神经元，该神经元所属类别即为此数据的聚类标签。K-means 算法聚类 Iris 数据集时，设定 k 值为 3，经过多次迭代形成 3 个聚类簇，且簇中心节点的特征向量表征了该簇数据所属类别。混合 SOM 算法首先训练 SOM 模型，拟合数据后对过于稀疏或低于阈值的节点进行剔除，在 SOM 模型之后增加一层 KNN 模型，基于质心与观测数据点之间的欧氏距离对数据进行分类。PCA-SOM 算法先利用 PCA 将数据矩阵投影到较小的子空间中以降低原始数据维度；然后通过建立数据间的统计相关性，将更能代表数据特征的特性作为 SOM 模型的输入；最后 SOM 进一步根据数据间的欧氏距离将其分类。icSOM 模型使用 K-means 算法对数据进行预分类，分别训练得到属于不同类别数据的 SOM 模型，随后，测试数据在不同的 SOM 模型中寻找获胜神经元，并通过计算比较得到置信神经元，该神经元所属类别即为此数据的聚类标签。

表 1　不同算法聚类 Iris 效果对比

Table 1　Comparison of clustering effects on Iris dataset by different algorithms

算法	*S*	*ARI*	*P*	*R*	*F*1	*ACC*	*AOT*/ms
传统 SOM	0.5109	0.9583	0.88	0.85	0.86	0.8986	0.0318
K-means	0.5221	0.9610	0.90	0.88	0.89	0.9093	0.0245
混合 SOM	0.5143	0.9724	0.93	0.90	0.91	0.9268	0.0675
PCA-SOM	0.5127	0.9602	0.91	0.88	0.89	0.8999	0.0591
icSOM	0.5262	0.9697	0.93	0.92	0.92	0.9333	0.2217

由表 1 可知，icSOM 算法的 *S* 值最高，为 0.5262，高于传统的 SOM 和 K-means 算法，表明经过 K-means 算法预分类后的改进 SOM 模型可以实现更高的簇内紧密度以及簇间疏离度。引入 Iris 数据集的真实标签用于外部评价指标分析。从广义上说，ARI 衡量的是两个数据的吻合程度。表 1 中混合 SOM 算法的 *ARI* 值最高，为 0.9724，这是由于在其用 SOM 模型对数据去噪时，舍弃掉了一部分不符合阈值设定的神经元节点以及映射到这些节点的样本数据，式(11)中的 C_2^n 值相比其他算法发生了变化，有可能影响最终 *ARI* 值的计算。icSOM 算法的 *ARI* 值次之，为 0.9697，也十分接近 1。传统 SOM 模型的 *F*1 值为 0.86，*ACC* 值为 0.8986。先利用 PCA 算法将数据投影到低维子空间，消除各特征向量间的相互影响，再将更能表征数据特点的向量作为 SOM 模型的输入，则可将 *F*1 指标提升至 0.89，*ACC* 值也略有提高。将传统 SOM 模型与 KNN 算法结合后改进为混合 SOM 模型后，上述两个指标值均有所提高，分别为 0.91 和 0.9268，说明通过阈值判断除去部分数据信息以提高数据集有效性的方法是可行的，但可能会存在因损失部分有效信息而产生聚类误差的情况。icSOM 算法的 *F*1 值为 0.92，*ACC* 值为 0.9333，这表明根据本文算法所提出的数据预分类，综合判断神经元的置信度及其与数据间的欧氏距离进而得到置信神经元，并据此给数据分配聚类标签的过程分析数据可以取得较好的聚类效果。PCA-SOM 算法通过对数据降维，减小计算开销。混合 SOM 算法通过阈值筛选，减少输入输出数量，降低了算法复杂度。icSOM 算法分别训练得到属于不同类别数据的权重矩阵，但测试数据需要在各类别权重矩阵中寻找获胜神经元并进一步分析判别，增加了时间开销，因此算法的 *AOT* 值较大，但依然符合数据实时处理要求。

由表 2 可得，相较于处理 Iris，在聚类 Wine 时，与 SOM 模型相关的各算法的 *F*1 值与 *ACC* 值均略有降低，*AOT* 值有所增加。这是 Wine 数据集的特性导致的，各类别样本数据数量分布不均匀可能会使得 SOM 模型在训练过程中各类别产生的影响不均匀，更多的特征向量组成高维数据，增加了算法运算时间。K-means 算法在两个数据集上的效果基本一致，说明该算法对于高维数据有较好的可扩展性，且收敛速度快、算法时间复杂度较低。传统 SOM 算法的 *ACC* 值下降程度大，原因在于其仅对数据进行 SOM 聚类，各个特征间彼此影响且影响程度受数量分布不均而不同，增大了聚类误差。icSOM 算法中，数据先经过 K-means 算法预分类，再分别用于训练属于不同类的 SOM 模型，最后通过置信神经元判断数据类别，减小了因样本分布特性而产生的聚类误差，*ACC* 值仅有轻微下降。PCA-SOM 算法的 *AOT* 值基本不变，其通过降维处理控制 SOM 模型输入数据的维度。icSOM 在处理 Wine 时，*AOT* 值增大为 0.2262 ms，尽管符合数据实时处理要求，但也反映出该算法在处理大规模复杂数据时有一定局限性。

表 2　不同算法聚类 Wine 效果对比

Table 2　Comparison of clustering effects on Wine dataset by different algorithms

算法	*S*	*ARI*	*P*	*R*	*F*1	*ACC*	*AOT*/ms
传统 SOM	0.5084	0.9570	0.87	0.85	0.86	0.8873	0.0321
K-means	0.5219	0.9598	0.90	0.88	0.89	0.9091	0.0248
混合 SOM	0.5145	0.9721	0.93	0.91	0.92	0.9266	0.0680
PCA-SOM	0.5122	0.9594	0.90	0.88	0.89	0.8961	0.0597
icSOM	0.5261	0.9697	0.92	0.91	0.91	0.9298	0.2262

结束语　本文在经典无监督学习 SOM 模型的基础上进行创新，提出了 icSOM 模型，可用于对数据进行聚类分析。样本数据先由 K-means 算法进行初步分类，然后将预分类后的数据分别用于训练以得到不同的 SOM 模型，在提出置信度矩阵概念后，通过综合判断获胜神经元的置信度及其与输入数据间的欧氏距离得到置信神经元，根据置信神经元所属类别给数据分配聚类标签。该算法通过对数据进行预分类，为传统无监督学习提供了更多的数据信息量；分类训练 SOM 模型在一定程度上消除了不同类之间的影响；对获胜神经元的综合判断方法使得数据集中的模糊点更易被聚类。本文以 Iris 和 Wine 数据集为例，依据各内部、外部评价指标对 icSOM 算法性能进行分析，实验结果表明，该算法可以更好地处理样本数据且有较好的聚类效果，但其处理大规模复杂数据时在算法复杂度方面存在一定局限性。未来可以将 icSOM 算法与降维算法相结合，以进一步降低处理数据的算法复杂度。

参 考 文 献

[1] SCHANK R C. What is AI, anyway? [J]. AI Magazine, 1987, 8(4): 59.

[2] MCCARTHY J. Generality in artificial intelligence [J]. Communications of the ACM, 1987, 30(12): 1030-1035.

[3] SAMUEL A L. Some studies in machine learning using the game of checkers [J]. IBM Journal of Research and Development, 1959, 3(3): 210-229.

[4] SAMUEL A L. Machine learning [J]. The Technology Review, 1959, 62(1): 42-45.

[5] BA ŞTANLAR Y, ÖZUYSAL M. Introduction to machine learning [J]. miRNomics: MicroRNA Biology and Computational Analysis, 2014: 105-128.

[6] EL NAQA I, MURPHY M J. What is machine learning? [M]. Springer International Publishing, 2015.

[7] ZHOU Z H. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.

[8] NASTESKI V. An overview of the supervised machine learning methods [J]. Horizons. B, 2017, 4: 51-62.

[9] DAYAN P, SAHANI M, DEBACK G. Unsupervised learning [J]. The MIT Encyclopedia of the Cognitive Sciences, 1999: 857-859.

[10] BZDOK D, KRZYWINSKI M, ALTMAN N. Machine learning: supervised methods [J]. Nature Methods, 2018, 15(1): 5-6.
[11] LAAKSONEN J, OJA E. Classification with learning k-nearest neighbors [C]// Proceedings of International Conference on Neural Networks. IEEE, 1996: 1480-1483.
[12] LIU C X, SHI D M, SONG W J. Research thread and latest progress of the methods of dimensionality reduction in high-dimensional data[J]. Journal of Statistics, 2023, 4(3): 11-21.
[13] WOLD S, ESBENSEN K, GELADI P. Principal component analysis [J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2(1/2/3): 37-52.
[14] BAI Y X. The application of k-means in feature selection[J]. Electronic Technology & Software Engineering, 2018, 123(1): 186-187.
[15] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [J]. KDD, 1996, 96(34): 226-231.
[16] KOHONEN T. The self-organizing map [J]. Proceedings of the IEEE, 1990, 78(9): 1464-1480.
[17] KOHONEN T. Things you haven't heard about the self-organizing map [C]// IEEE International Conference on Neural Networks. IEEE, 1993: 1147-1156.
[18] KOHONEN T. Exploration of very large databases by self-organizing maps [C]// Proceedings of International Conference on Neural Networks. IEEE, 1997, 1: 1-6.
[19] KOHONEN T. Essentials of the self-organizing map [J]. Neural Networks, 2013, 37: 52-65.
[20] ZHOU G, YANG F, XIAO J. Study on pixel entanglement theory for imagery classification [J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-18.
[21] LI S, LIU F, JIAO L, et al. Self-supervised self-organizing clustering network: a novel unsupervised representation learning method [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 35: 1857-1871.
[22] YAN J, ZHANG C, LI Y. A clustering method for power time series curves based on improved self-organizing map algorithm [C]// 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI). IEEE, 2023: 451-455.
[23] XIE D, FAN L, FU C, et al. Nonintrusive load monitoring algorithm using SOM-AdaDBSCAN [C]// 2023 6th International Conference on Energy, Electrical and Power Engineering (CEEPE). IEEE, 2023: 905-910.
[24] KHAN S, MAILEWA A B. Discover botnets in IoT sensor networks: A lightweight deep learning framework with hybrid self-organizing maps [J]. Microprocessors and Microsystems, 2023, 97: 104753.
[25] BENDJAMA H, BOUHOUCHE S, AOUABDI S, et al. Monitoring of casting quality using principal component analysis and self-organizing map[J]. The International Journal of Advanced Manufacturing Technology, 2022, 120(5): 3599-3607.
[26] FORT J C, PAGÈS G. About the Kohonen algorithm: strong or weak self-organization? [J]. Neural Networks, 1996, 9(5): 773-785.
[27] ANDERSON E. The irises of the gaspe peninsula [J]. Bulletin of American Iris Society, 1935, 59: 2-5.
[28] AEBERHARD S, COOMANS D, VEL O D. Comparative analysis of statistical pattern recognition methods in high dimensional settings [J]. Pattern Recognition, 1994, 27(8): 1065-1077.

JIANG Rui, born in 1985, Ph. D, associate professor. His main research interests include artificial intelligence and wireless communication.

(责任编辑:何杨)