基于堆叠稀疏自编码的模糊C-均值聚类算法

段宝彬[1,2],韩立新[2],谢 进[1]
DUAN Baobin[1,2], HAN Lixin[2], XIE Jin[1]

1.合肥学院 数学与物理系,合肥 230601
2.河海大学 计算机与信息学院,南京 211100
1.Department of Mathematics and Physics, Hefei University, Hefei 230601, China
2.College of Computer and Information, Hohai University, Nanjing 211100, China

DUAN Baobin, HAN Lixin, XIE Jin. Fuzzy C-means clustering algorithm based on stacked sparse autoencoders. Computer Engineering and Applications, 2015, 51(4):154-157.

Abstract: In order to solve the sensitivity of fuzzy C-means clustering algorithm to the outlier and the randomly initialized clustering center, the stacked sparse autoencoders and traditional fuzzy C-means clustering algorithm are combined to improve the traditional fuzzy C-means clustering algorithm. Because the stacked sparse autoencoders can extract features of the original data set from low-level to high-level, and high-level features can reflect the nature features of the sample data to be clustered better than the original data set, which will help to improve the clustering effect with high-level features instead of the original data. With experimenting on several standard data sets of UCI, it is shown that the improved algorithm is feasible.
Key words: stacked sparse autoencoders; fuzzy C-means clustering; features; deep learning

摘 要:针对模糊C-均值聚类算法对孤立点、随机初始化的聚类中心比较敏感的问题,将堆叠稀疏自编码与传统模糊C-均值聚类算法相结合,对传统模糊C-均值聚类算法进行了改进。由于堆叠稀疏自编码可以提取原始数据集从低层到高层的特征,而高层的特征通常比原始数据集更能反映待聚类样本的本质特征,用其代替原始数据集进行聚类,有助于提高聚类的效果。利用改进后的算法在UCI的几个标准数据集上进行实验,结果表明改进后的算法是有效可行的。
关键词:堆叠稀疏自编码;模糊C-均值聚类;特征;深度学习
文献标志码:A **中图分类号:**TP301 doi:10.3778/j.issn.1002-8331.1402-0149

1 引言

近年来,一种在图像分类、语音识别、自然语言处理等领域获得巨大成功的机器学习技术日益受到工业界、学术界的广泛关注,这就是深度学习技术。微软、谷歌、百度等著名高科技公司开始投入大量资金和人力用来支持深度学习技术的研发。深度学习通过构建含多个隐层的机器学习模型,输入海量的训练数据,以学习原始数据更本质的特征,有助于最终提高预测或决策的正确率[1]。常用的深度学习模型有堆叠稀疏自编码[2]、深度信念神经网络[3]和深度卷积神经网络[4]等,其中堆叠稀疏自编码是一种最容易实现的深度学习模型。

模糊C-均值聚类(FCM)是目前应用最广泛的一种软聚类算法,它基于划分方法解决聚类问题,已成功应用于数据挖掘、模式识别等领域。该算法是由Dunn[5]于1973年提出的,Bezdek[6] 于1981年进一步完善了该算法。模糊C-均值聚类的基本思想是首先根据给定的聚类数随机初始化聚类中心,然后采用迭代方法不断计算各样本隶属于各类别的模糊隶属度,按最大模糊隶属度

基金项目:江苏省高校“青蓝工程”中青年学术带头人培养对象资助项目;安徽省自然科学基金项目(No.1208085MA15);合肥学院应用数学重点建设学科基金(No.2014xk08)。
作者简介:段宝彬(1975—),男,博士研究生,副教授,研究领域为机器学习,数据挖掘;韩立新(1967—),男,博士生导师,教授,研究领域为Web技术,信息检索,模式识别,数据挖掘;谢进(1970—),男,博士,副教授,研究领域为计算机辅助几何设计及计算机图形学。E-mail:duanbb@126.com
收稿日期:2014-02-18 **修回日期:**2014-05-27 **文章编号:**1002-8331(2015)04-0154-04
CNKI**网络优先出版:**2014-07-11, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1402-0149.html

原则调整样本类别并重新计算聚类中心，从而对目标函数进行优化，以实现对样本数据集的自动聚类[7]。但传统模糊C-均值聚类算法直接采用原始数据集进行训练，对样本中的孤立点及随机初始化的聚类中心比较敏感，优化时容易使目标函数陷入局部极小[8]。为了克服这些缺点，不少学者对传统FCM算法进行了改进。Li[9]利用非参数加权特征提取的加权均值代替随机初始化的聚类中心，在一些数据集上得到了较好的结果；蔡静颖[10]采用马氏距离代替传统FCM算法的欧式距离，并在目标函数中引入协方差矩阵的调节因子，在一些数据聚类和图像分割实验中获得了满意的结果。伍忠东等[11]将核方法应用到模糊C-均值聚类中，提出了基于核函数的模糊核C-均值算法，使其能够解决非超球体数据、多种模式混合数据、含噪声数据、不对称数据的聚类问题。由于堆叠稀疏自编码通过训练能够提取输入数据从低层到高层的本质特征，而且提取到的特征对输入数据的畸变具有一定的容忍度，因此，本文尝试对传统模糊C-均值聚类算法进行改进，利用堆叠稀疏自编码提取原始数据集的高层特征，这些高层特征通常更能刻画样本数据的本质特性，在一定程度上可以减少聚类中心初始值和噪音等对聚类结果的影响，有助于提高聚类的效果。最后利用改进后的算法在UCI的四个标准数据集上进行实验，结果表明，本文提出的基于堆叠稀疏自编码的模糊C-均值聚类算法（以下简记为SAEFCM算法）是有效可行的，提高了聚类结果的正确率。

2　堆叠稀疏自编码

2.1　稀疏自编码

一个典型的自编码结构[12-13]如图1所示。

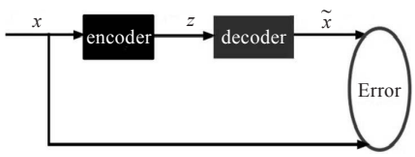

图1　自编码结构

假设原始数据全部为无标签数据 $\boldsymbol{x}\in\boldsymbol{R}^{d\times 1}$，将 $\boldsymbol{x}$ 输入到具有 N 个神经元的编码层，通过一个非线性的激活函数 f，如Sigmoid函数等，由公式(1)得到相应的映射 $\boldsymbol{z}\in\boldsymbol{R}^{N\times 1}$。

$$\boldsymbol{z}=f(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}_1) \tag{1}$$

其中 $\boldsymbol{W}\in\boldsymbol{R}^{N\times d}$ 是权重矩阵，$\boldsymbol{b}_1\in\boldsymbol{R}^{N\times 1}$ 是编码层偏置向量。

然后将 $\boldsymbol{z}$ 输入到解码层，由公式(2)得到对原始数据重构后的映射 $\tilde{\boldsymbol{x}}\in\boldsymbol{R}^{d\times 1}$。

$$\tilde{\boldsymbol{x}}=f(\tilde{\boldsymbol{W}}\boldsymbol{z}+\boldsymbol{b}_2) \tag{2}$$

为减少需要训练的参数个数，一般可取 $\tilde{\boldsymbol{W}}=\boldsymbol{W}^{\mathrm{T}}$，其中 $\boldsymbol{W}^{\mathrm{T}}$ 为权重矩阵 $\boldsymbol{W}$ 的转置，$\boldsymbol{b}_2\in\boldsymbol{R}^{d\times 1}$ 是解码层偏置向量。

通过训练调整权重矩阵 $\boldsymbol{W}$ 和偏置向量 $\boldsymbol{b}_1,\boldsymbol{b}_2$，使重构误差

$$L(\boldsymbol{x},\tilde{\boldsymbol{x}})=\frac{1}{2}\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2$$

达到极小。

若在自编码的基础上增加稀疏性约束条件，使编码层的输出 $\boldsymbol{z}$ 大部分元素为0，只有少数不为0，这就是稀疏自编码。此时，稀疏自编码重构误差函数为：

$$L_s(\boldsymbol{x},\tilde{\boldsymbol{x}})=\frac{1}{2}\|\boldsymbol{x}-\tilde{\boldsymbol{x}}\|^2+\beta\sum_{j=1}^{N}\left[\rho\ln\frac{\rho}{\hat{\rho}_j}+(1-\rho)\ln\frac{1-\rho}{1-\hat{\rho}_j}\right] \tag{3}$$

其中 β 是稀疏性惩罚因子的权重；ρ 为稀疏性参数，一般取接近于0的正数；$\hat{\rho}_j$ 表示第 j 个特征在整个训练数据集上的平均激活度，它间接取决于权重矩阵 $\boldsymbol{W}$ 和偏置向量 $\boldsymbol{b}_1$。

输入训练数据集，采用BP算法[14]和L-BFGS优化算法[15]进行训练，调整权重矩阵 $\boldsymbol{W}$ 和偏置向量 $\boldsymbol{b}_1$、$\boldsymbol{b}_2$，使式(3)达到极小。

2.2　堆叠稀疏自编码结构

如果将第一个稀疏自编码层的输出 $\boldsymbol{z}$ 作为第二个稀疏自编码层的输入，同样利用极小化重构误差原则进行训练，得到第二个稀疏自编码层的输出，以此类推，将第 $n-1$ 个稀疏自编码层的输出作为第 n 个稀疏自编码层的输入，逐层利用贪婪算法进行训练，可得到堆叠稀疏自编码，如图2所示。

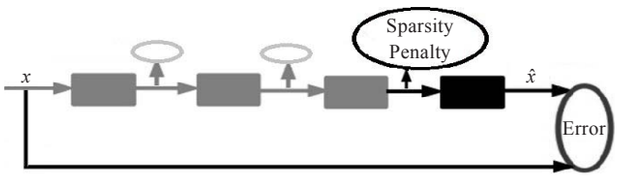

图2　堆叠稀疏自编码结构

3　模糊C-均值聚类算法及其改进

3.1　传统模糊C-均值聚类算法

设原始无标签数据集为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中 $\boldsymbol{x}_i\in\boldsymbol{R}^{d\times 1}$，$i=1,2,\cdots,n$；根据模糊隶属度可将 $\boldsymbol{X}$ 分为 c 个子类 $S_1,S_2,\cdots,S_c$，相应的 c 个聚类中心分别为 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c$；用 u_{ij} 表示第 i 个样本数据 $\boldsymbol{x}_i$ 隶属于第 j 个子类 S_j 的隶属度，m 是用来控制聚类结果的模糊权重，要求 $m>1$，一般可取2；采用迭代算法极小化如下目标函数：

$$J_m=\sum_{j=1}^{c}\sum_{i=1}^{n}u_{ij}^{m}\|\boldsymbol{x}_i-\boldsymbol{v}_j\|^2 \tag{4}$$

其中 u_{ij} 满足如下约束条件：

$$0\leqslant u_{ij}\leqslant 1,\ i=1,2,\cdots,n;j=1,2,\cdots,c$$

$$\sum_{j=1}^{c}u_{ij}=1,\ i=1,2,\cdots,n$$

$$\sum_{i=1}^{n}u_{ij}>0,\ j=1,2,\cdots,c$$

利用拉格朗日乘子法易得，当 J_m 取极小值时，对应的模糊隶属度和聚类中心分别如下：

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{v}_j\|}{\|\boldsymbol{x}_i-\boldsymbol{v}_k\|}\right)^{\frac{2}{m-1}}},i=1,2,\cdots,n;j=1,2,\cdots,c \quad (5)$$

$$v_j=\frac{\sum_{i=1}^{n}u_{ij}^m\boldsymbol{x}_i}{\sum_{i=1}^{n}u_{ij}^m},j=1,2,\cdots,c \quad (6)$$

具体实现时，首先随机初始化聚类中心 $v_j(j=1,2,\cdots,c)$，然后利用原始样本数据和公式(5)、(6)通过迭代算法对隶属度 $u_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,c)$ 和聚类中心 $v_j(j=1,2,\cdots,c)$ 进行更新。

3.2　改进的模糊C-均值聚类学习算法

3.2.1　ZCA白化

在使用聚类算法之前，为了减少原始数据的冗余性，需要对数据进行白化处理[16]，使得所有属性的方差相同，不同属性之间不相关或具有较低的相关性。常用的一种白化方法是ZCA白化，它可以使得白化后的数据尽可能接近原始数据，并且保持与原始数据相同的维数。设原始数据集有 n 个样本，每个样本的维数为 d，则ZCA白化的过程如下：

(1)首先将原始数据集排成一个 $d\times n$ 的数值矩阵 $\boldsymbol{X}$，然后进行使每个属性均值为零的标准化处理，得到的矩阵记为 $\boldsymbol{A}$。

(2)计算 $\boldsymbol{A}$ 对应的样本协方差矩阵 $\boldsymbol{\Sigma}$，求出相应的特征值，并按从大到小顺序分别记为 $\lambda_1,\lambda_2,\cdots,\lambda_d$，对应的特征向量分别记为 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_d$，并记 $\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_d]$。

(3)计算旋转后的矩阵

$$\boldsymbol{X}_{\text{rot}}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X}=\begin{bmatrix}\boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{u}_1^{\mathrm{T}}\boldsymbol{x}_n\\ \boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{u}_2^{\mathrm{T}}\boldsymbol{x}_n\\ \vdots & \vdots & & \vdots\\ \boldsymbol{u}_d^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{u}_d^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{u}_d^{\mathrm{T}}\boldsymbol{x}_n\end{bmatrix}$$

为使旋转后矩阵对应的每个属性具有单位方差，可分别用 $1/\sqrt{\lambda_i}\ (i=1,2,\cdots,d)$ 去乘以矩阵 $\boldsymbol{X}_{\text{rot}}$ 相应第 $i(i=1,2,\cdots,d)$ 行的各元素，得到的矩阵记为 $\boldsymbol{X}_{\text{rot}1}$。

(4)将矩阵 $\boldsymbol{X}_{\text{rot}1}$ 左乘矩阵 $\boldsymbol{U}$，则得到的矩阵 $\boldsymbol{X}_1=\boldsymbol{U}\boldsymbol{X}_{\text{rot}1}$ 就是原始数据集ZCA白化的结果，矩阵的每一列对应ZCA白化后的样本数据。

在ZCA白化过程中，若存在 λ_i 接近于0，则可在上述第3步中用 $1/\sqrt{\lambda_i+\varepsilon}$（$\varepsilon$ 可取一个很小的正的常数，如0.1）代替 $1/\sqrt{\lambda_i}$，以避免出现数值不稳定或数据上溢的现象。

3.2.2　SAEFCM算法的主要步骤

以含有两个编码层的堆叠稀疏自编码为例，本文所提出的SAEFCM算法的主要步骤如下：

步骤1　采用3.2.1中ZCA白化方法对原始输入数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$ 进行预处理，得到 $\boldsymbol{X}$ 的低冗余性表示 $\boldsymbol{X}_1$。

步骤2　用 $\boldsymbol{X}_1$ 训练堆叠稀疏自编码器的第一个编码层，利用极小化重构误差原则调整权重矩阵 $\boldsymbol{W}_1$ 和偏置矩阵 $\boldsymbol{B}_1$，由公式(7)得到数据集的第一层特征表示 $\boldsymbol{Z}_1$。

$$\boldsymbol{Z}_1=f(\boldsymbol{W}_1\boldsymbol{X}_1+\boldsymbol{B}_1) \quad (7)$$

步骤3　将第一个稀疏编码层的输出 $\boldsymbol{Z}_1$ 输入到第二个稀疏编码层，同样利用极小化重构误差原则调整权重矩阵 $\boldsymbol{W}_2$ 和偏置矩阵 $\boldsymbol{B}_2$，由公式(8)得到数据集的第二层特征表示 $\boldsymbol{Z}_2$。

$$\boldsymbol{Z}_2=f(\boldsymbol{W}_2\boldsymbol{Z}_1+\boldsymbol{B}_2) \quad (8)$$

步骤4　根据文献[17]的方法确定最佳聚类数，将第二个稀疏自编码层的输出 $\boldsymbol{Z}_2$ 转置后代替原始数据集，最后利用3.1中传统模糊C-均值聚类算法得到相应的聚类结果。

3.3　相关工作的比较

传统模糊C-均值聚类算法和本文提出的改进模糊C-均值聚类算法的后半部分是完全相同的，都是先随机初始化聚类中心，利用迭代算法更新模糊隶属度和聚类中心。

传统模糊C-均值聚类算法和本文提出的改进模糊C-均值聚类算法的不同点是传统模糊C-均值聚类算法直接利用无标签原始数据集进行模糊隶属度和聚类中心的计算和更新；而本文提出的改进模糊C-均值聚类算法先采用ZCA白化方法对原始数据集进行预处理，消除数据之间的冗余性；然后利用堆叠稀疏自编码，逐层提取原始数据的各级特征，利用更能反映样本数据本质属性的高层特征进行模糊聚类分析，有助于提高聚类结果的准确性和鲁棒性。

4　实验及结果分析

本实验是基于Win7 64位操作系统，CPU为Intel I5-2450M，2.5 GHz，内存为6 GB，所用软件为Matlab 2013a。从UCI机器学习数据集中选择四个常用标准数据集来测试本文提出的改进模糊C-均值聚类算法。实验采用含有两个编码层的稀疏自编码结构，同一般神经网络隐层神经元数目一样，两个自编码层神经元的数目 N_1 和 N_2 的确定也没有统一的方法，根据具体数据集的属性数、样本数、类别数等通过实验确定合适的值，如对于数据集Iris，N_1 和 N_2 均取20；对于数据集Pima，N_1 和 N_2 分别取100和200。其他参数一般根据经验选取，本实验中取 $\beta=3$，$\rho=0.1$。

四个标准测试数据集的简单描述如表1所示。

将本文所提出的SAEFCM算法与传统模糊C-均值(FCM)、核模糊C-均值聚类(KFCM)算法在上述四个数据集上进行测试。为减少随机化权重、初始聚类中心等

表1　数据集描述

数据集名	样本数	属性数	类别数
Iris	150	4	3
Wine	178	13	3
Pima	768	8	2
Glass	214	9	6

对聚类结果的影响,每个聚类算法各运行20次,聚类的平均正确率比较结果如表2所示,运行时间比较结果如表3所示。

表2　三种模糊C-均值聚类算法的平均正确率　(%)

数据集名	SAEFCM	FCM	KFCM
Iris	94.90	89.33	89.83
Wine	88.40	68.54	71.94
Pima	69.79	65.89	67.29
Glass	61.33	60.75	60.98

表3　三种模糊C-均值聚类算法的运行时间比较　s

数据集名	SAEFCM	FCM	KFCM
Iris	46	0.36	0.18
Wine	185	0.55	0.48
Pima	739	0.62	0.78
Glass	202	0.65	0.55

从表2可以看出,本文所提出的SAEFCM算法在四个标准数据集上的性能均优于传统FCM算法和KFCM算法,尤其在Wine数据集上,聚类正确率提高将近20%;由于堆叠稀疏自编码提取的特征更能反映样本数据的本质属性,因此,提高了聚类的正确率。

从表3可以看出,本文所提出的SAEFCM算法的时间复杂度较高,在四个标准数据集上的运行时间均远大于传统FCM算法和KFCM算法,这是由于堆叠稀疏自编码特征学习时需要根据随机的初始权重利用极小化重构误差原则优化权重,相应的迭代计算量比较大,尤其当数据集属性数、类别数和自编码层神经元数比较大时运行时间更长,因此,SAEFCM算法虽然提高了聚类的正确率,但也增加了计算的时间复杂度,因此,本文提出的SAEFCM算法对于实时性要求较高的聚类问题不太适合。

5　结束语

本文对传统模糊C-均值聚类算法进行了改进,提出了基于堆叠稀疏自编码的模糊C-均值聚类(SAEFCM)算法,给出了算法实现的具体步骤,最后将其用于UCI四个常用标准数据集的聚类问题中,取得了较好的聚类效果。

由于软硬件限制,本文只采用了两个稀疏自编码层的结构进行实验,进一步加深自编码层数是否能显著提高聚类效果还有待进一步研究。另外,稀疏自编码中相关参数如何进一步优化、如何缩短特征的学习时间等问题都有待进一步研究。

参考文献:

[1] 余凯,贾磊,陈雨强,等.深度学习的昨天、今天和明天[J].计算机研究与发展,2013,59(12):1799-1804.

[2] Ranzato M,Poultney C,Chopra S,et al.Efficient learning of sparse representations with an energy-based model[C]//NIPS,2006.

[3] Hinton G E,Osindero S,Teh Y W.A fast learning algorithm for deep belief nets[J].Neural Computation,2006,18(7):1527-1554.

[4] Krizhevsky A,Sutskever I,Hinton G E.ImageNet classification with deep convolutional neural networks[C]//NIPS 2012:1106-1114.

[5] Dunn J C.A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J].Journal of Cybernetics,1973(3):32-57.

[6] Bezdek J C.Pattern recognition with fuzzy objective function algoritms[M].New York:Plenum Press,1981.

[7] 于剑.论模糊C均值算法的模糊指标[J].计算机学报,2003,26(8):968-973.

[8] Pham D T,Otri S,Afify A,et al.Data clustering using the bees algorithm[C]//Proceedings of the 40th CIRP International Seminar on Manufacturing Systems,May 30-June 1,2007,Liverpool,UK,2007.

[9] Li C H,Huang W C,Kuo B C,et al.A novel fuzzy weighted C-means method for image classification[J].Int J Fuzzy Syst,2008,10(3):168-173.

[10] 蔡静颖,谢福鼎,张永.基于自适应马氏距离的模糊C均值算法[J].计算机工程与应用,2010,46(34):174-176.

[11] 伍忠东,高新波,谢维信.基于核方法的模糊聚类算法[J].西安电子科技大学学报:自然科学版,2004,31(4):533-537.

[12] Zouxy.Deep Learning(深度学习)学习笔记整理系列之(四)[EB/OL].[2014-02-09].http://blog.csdn.net/zouxy09/article/details/8775524.

[13] Shin H C,Orton M R,Collins D J,et al.Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(8):1930-1943.

[14] Rumelhart D E,Hinton G E,Williams R J.Learning representa-tions by back-propagating errors[J].Nature,1986,323:533-536.

[15] Liu D C,Nocedal J.On the limited memory BFGS method for large scale optimization[J].Mathematical Program Ming,1989,45:503-528.

[16] Ng A,Ngiam J,Foo Chuan Yu,et al.Whitening[EB/OL].[2014-04-12].http://deeplearning.stanford.edu/wiki/index.php/Whitening.

[17] 周世兵.聚类分析中的最佳聚类数确定方法研究及应用[D].无锡:江南大学,2011.