一种新的基于深度学习的聚类分析算法

侯远韶

(河南工业贸易职业学院 信息工程系,河南 郑州 451191)

摘 要:传统图像分类算法需要大量的数据进行有监督训练,造成了数据冗余进而带来维数灾难,为此提出一种基于深度学习的聚类分析算法。聚类分析利用数据内部簇结构和模式进行分类,不需要对样本进行训练得到先验知识,降低了计算复杂度。引入深度学习对数据内部结构和模式进行特征学习,得到数据的初步聚类,再对初步聚类进行不断优化得到最终的分类效果。实验结果表明,算法很好地解决了信息全面与维数灾难的矛盾,具有良好的实用性和主观一致性。

关键词:深度学习;特征提取;聚类分析;无监督学习

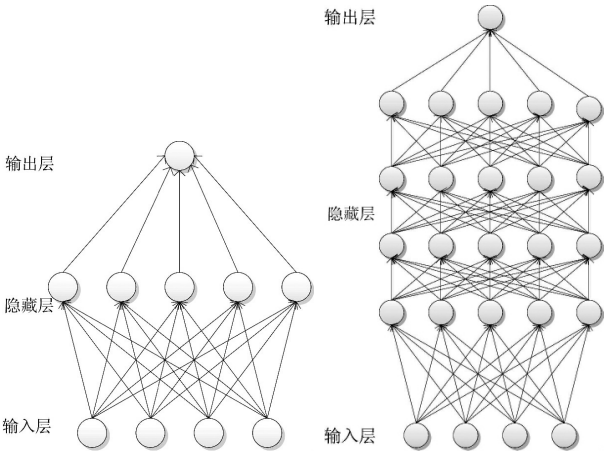
中图分类号:TP391.4 文献标识码:A 文章编号:2095-7726(2018)12-0021-04

图像分类是图像处理中重要的一环,它根据图像信息中含有的不同特征,把不同类别的数据区分开,从而得到图像分类效果[1]。目前图像分类的方法主要有:基于色彩特征的索引技术、基于纹理的图像分类技术、基于形状的图像分类技术和基于空间的图像分类技术。但当前的图像分类技术过于依赖图像特征的提取,算法学习大多为浅层模型,同时需要大量的训练数据进行样本训练(即进行有监督训练),这导致数据量增大和计算复杂度提高。深度学习的出现解决了这一难题。深度学习减小了传统训练算法的局部最小性,利用底层特征合成抽象的高层来表示数据的属性特征,不需要人为设计特征提取模型,在不改变数据原始信息的情况下降低数据维度使数据处理更加容易。聚类技术是数据挖掘的一种重要方法,是无监督学习方法,它通过分析数据集中包含的相似元素集合即簇结构来得到图像的分类[2]。将深度学习和聚类技术结合起来,可以更好地解决现实中的问题。

1 深度学习理论

1.1 深度学习的思想

深度学习的"深度"是相对于稀疏编码、最大熵及马尔可夫模型(HMM)等传统浅层机器学习算法而言的。这些浅层学习算法只能处理简单的数据,由于数据的输入层和输出层之间只有一层隐含节点,在面对复杂的非线性问题时,难以得到令人满意的结果。深度学习通过生成性训练避免过拟合现象的发生,将原始样本空间中的特征表示进行映射变换得到新的特征空间[3]。深度学习强调网络的深度,将原本需要多个函数多层次表示的问题,通过利用较少的参数将构造复杂的函数表示出来。深度学习每层的输出都是下一层的输入,输入层和输出层之间构造多个含有节点的隐藏层,通过分析层与层之间的线性组合关系来对其参数进行优化,得到图像分类和回归问题的解决方法[4]。浅层学习模型和深层学习模型如图 1 所示。



(a)浅层学习模型 (b)深层学习模型

图 1 浅层学习模型和深层学习模型

1.2 常见的深度学习模型

深度学习是一种无监督学习算法,它有多个隐藏层,而这若干个隐藏层可以看作是一个层,故在整体上

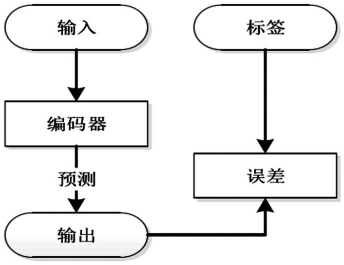
收稿日期:2018-10-20

基金项目:河南省科技攻关计划项目(0721002210032)

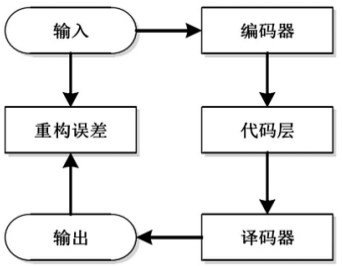
作者简介:侯远韶(1986—),男,河南平顶山人,讲师,硕士,研究方向:机器视觉与图像处理。

深度学习可看作一种分层计算的学习模型[5]。深度学习模型的主要步骤为：1)为了得到各层的数据参数，利用无监督学习方法对网络层进行由下至上的逐层训练。这一部分称为预训练，主要通过受限玻尔兹曼机(RBM)完成。2)将底层的训练结果作为下一层网络的输入进行再训练。3)对带有标签的数据进行有监督学习，以此对得到的结果进行微调进而消除误差[6]。有监督学习是示例学习；而无监督学习是观察学习，目标是寻找群体相似性。有监督学习和无监督学习流程如图2所示。

总的来说，深度学习是具有多层非线性映射的深层结构，是机器学习的一种模型。常见的深度学习模型可以分为：分类型深度结构，如深度卷积神经网(CNN)；合成型深度结构，如稀疏自动编码器、自动编码机等；生成型深度结构，如有限制玻尔兹曼机、PCANet和深度置信网络等。



(a)有监督学习



(b)无监督学习

图2 有监督学习和无监督学习流程

2 聚类分析

聚类技术是无监督学习，它通过分析数据集中包含的相似元素集合(即簇结构)来得到图像的分类，不同类型的子集之间要有尽可能大的差别，同一个类型子集内数据具有一致性[7]。聚类原理如图3所示。

聚类的数学定义为：假设 D 维空间中有 N 个数据结构，选取合理的阈值将 N 个数据结构划分为 K 个簇；阈值的选取的要求是不同类型子集之间的差别尽可能大，同一个类型子集内数据具有一致性。即数据集 X 中含有 N 个样本 $X_1, X_2, \cdots, X_N$，可划分为 K 个簇 $x_1, x_2, \cdots, x_K$，数据集和簇之间符合

$$\begin{cases} x_1 \cup x_2 \cup \cdots \cup x_K = X, \\ x_i \cap x_j = \varnothing (1 \le i \ne j \le K)。 \end{cases} \tag{1}$$

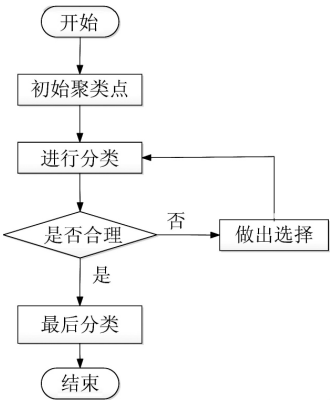


图3 聚类原理

聚类思想可以概括为：1)输入参数 K 作为划分簇的单元格依据；2)通过求解单元格中的类内数据来划分单元格类型；3)单元格之间和单元格内的属性划分通过求解熵 $I(x_j) \geqslant \log_2 k - I^*$ 和 $I(x_j) \leqslant I^*$ 得到，其中 x_j 为单元格划分属性依据，I^* 为输入参数；4)通过对簇的边界点和孤立点进行连接，得到不同类的簇。

聚类分析的类间和类内评判是算法成败的关键，要对数据进行合理分类，就必须对样本间的关系进行评判，而评判的依据主要是样本间的相似系数和距离[8]。

样本间相似系数接近1，表明它们属于同一类，可以划分为同一簇；相似系数接近0，表明它们属于不同类型，可以划分为不同簇。相似系数可以通过相关系数

$$C(x,y) = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_i (x_i - \overline{x})^2][\sum_i (y_i - \overline{y})^2]}} \tag{2}$$

和夹角余弦

$$H(x,y) = \frac{\sum_i x_i y_i}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}} \tag{3}$$

来表示。

样本之间的距离可以用来描述样本之间的亲疏程度。假设样本可以通过 N 个特征向量来表示，即样本仅仅是 N 维空间中的一个点，则样本之间距离近就可以判断为一个类别，距离远表示类别不同。对一个均值

为 μ,协方差矩阵为 Σ 的变量 x,其马氏距离表示为

$$D(x)=\sqrt{(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)}, \tag{4}$$

兰氏距离表示为

$$D(x,y)=\sum_i\frac{|x_i-y_i|}{|x_i+y_i|}, \tag{5}$$

明氏距离表示为

$$D_q(x,y)=(\sum|x_i-y_i|^q)^{\frac{1}{q}}。 \tag{6}$$

当式(6)中 q 分别取 1、2 和无穷大时,分别得到的是绝对距离、欧式距离和切比雪夫距离。

3 基于深度学习的聚类分析

传统的图像分类方法在处理高维数据时不仅无法避免维数灾难,还忽略了图像的结构信息,没有考虑图像的多个特征对象之间的关系,不能很好地描述数据的深层特征,而且单纯的聚类分析随着聚类过程中数据规模的增大,其隐形结构也越来越复杂,导致复杂的数据无法划分为简单的簇结构[9]。因此,如何将复杂问题简单化,用有限的特征数据划分事物的类型,需要进一步研究。

深度学习利用深层结构抽取事物的特征信息。通过深度学习思想提取聚类过程中的簇结构,可在避免过拟合的同时充分发挥深度学习和聚类技术的优点。聚类分析和深度学习具有很多相似点,如它们都是通过对数据内部模式和结构进行挖掘,都面临着大数据和维数灾难的挑战,而且同为无监督学习[10]。因此,本文提出一种新的基于深度学习的聚类分析算法,算法模型主要由预训练和微调两部分组成。算法思想为:1)对原始数据特征进行深度学习,利用相似系数和距离判别判定初始聚类;2)通过模糊判别得到样本的类别信息,即聚类中心的模糊位置;3)对聚类中心的模糊位置进行微调,即对样本间的相似系数和距离进行分析并优化其参数,得到最终的分类效果。

对判定的初始聚类进行微调,主要是采用交叉迭代的方法对聚类中心的模糊位置目标

$$\begin{aligned}f=e\times\{&-\sum_i x_i\log_2 f_i(R(x_i,w_i))-\\&\sum_i(1-x_i)\log_2[1-f_i(R(x_i,w_i))]\}+\\&(1-e)\times\sum_{i=1}^{N}\sum_{j=1}^{C}\sqrt{[R(x_i,w_i)-c_j]^2}\end{aligned} \tag{7}$$

进行微调[11]。式(7)中,$R(\cdot)$为深度学习对原始数据进行特征学习得到的新特征,e 为调节因子,$f(R(\cdot))$为解码数据即原始数据的重构,w 为参数隶属度矩阵,w_i 为参数隶属度矩阵中的元素,c_j 为单个元素的相关系数。

由于式(7)是收敛函数,可对其进行优化使目标函数更容易计算。优化后的目标函数为

$$\begin{aligned}f=e\times\{&-\sum_i x_i\log_2 f_i(R(x_i,w_i))-\\&\sum_i(1-x_i)\log_2[1-f_i(R(x_i,w_i))]\}+\\&\frac{1}{2}(1-e)\times\sum_{i=1}^{N}\sum_{j=1}^{C}[R(x_i,w_i)-c_j]^2。\end{aligned} \tag{8}$$

整个训练过程不需要对原始数据进行太多预处理,大大降低了计算复杂度,且对输入数据也具有很好的泛化性。模型可以根据实际需要应用于高维数据和低维数据处理中,灵活多变,能够充分显示数据本身存在的类别信息。

4 实验仿真

实验硬件环境为 Windows7 操作系统,CPU 为 Intel I7 处理器,16 G 内存,12 核 GPU。集成开发环境为 Microsoft Visual C++2010。聚类准确率采用 Gan 提出的评价标准:

$$r=\frac{\sum_{i=1}^{k}a_i}{n}, \tag{9}$$

其中 k 为聚类数,a_i 为分类的样本数,n 为样本数。聚类准确率 r 值越大分类效果越好,误差越小。r 值取值范围为 $0\leqslant r\leqslant 1$,取 0 时表示算法无法进行有意义聚类,取 1 时则表示聚类效果完全正确。但在一般情况下,r 值只能趋近于 1。

为了验证算法的有效性和实用性,采用具有代表性的 UCI 数据库作为实验对象。UCI 数据库是由加州大学提出的机器学习数据库,里面有不同数据集。本文通过对已有数据进行尺度变换、膨胀和腐蚀运算来扩充数据集大小。实验所用扩充后数据集参数如表 1 所示。

表 1 扩充后数据集信息

数据集名称	维数	数值型维数	类型属性	数据量
Aggregation	2	2	7	788
Jain	2	2	2	373
Spiral	2	2	3	312
Statlog	13	5	2	270

实验比较经典的 k-均值聚类算法、迭代最小平方误差聚类算法和本文基于深度学习的聚类分析算法在 UCI 数据库中的聚类准确率。具体结果如表 2 所示。由表 2 可知,在具有代表性的 UCI 数据库中,基于深度学

习的聚类分析算法聚类准确率明显优于其他三种算法。由于对数据集进行了尺度变换、膨胀和腐蚀运算，扩充了数据集样本容量，使样本数据具有不同的表现形式，说明该算法可以根据实际需要应用于不同数据集，灵活多变，具有很好的准确性和鲁棒性。

表 2　3 种算法在 UCI 数据库上的聚类准确率

算法	聚类准确率
k-均值聚类算法	0.716
迭代最小平方误差聚类	0.724
深度学习的聚类分析算法	0.825

5　结束语

笔者对深度学习理论进行阐述，分析了常见的深度学习模型以及聚类的定义，提出了将深度学习和聚类分析结合起来的算法。为了验证算法的有效性和适应性，通过在扩充后 UCI 数据库比较常规聚类分析算法和本文算法的聚类准确率，得出本文算法可以根据实际需要应用于不同数据集，灵活多变。但算法的时效性以及聚类分析相似度的定义有待进一步研究，而如何在降低计算量的同时寻找更加有效的算法是下一步研究的重点。

参考文献：

[1] 尹宝才，王文通，王立春.深度学习研究综述[J].北京工业大学学报，2015(1)：48–59.

[2] 孙志军，薛磊，许阳明，等.深度学习研究综述[J].计算机应用研究，2012，29(8)：2806–2810.

[3] 郑胤，陈权崎，章毓晋.深度学习及其在目标和行为识别中的新进展[J].中国图象图形学报，2014，19(2)：175–184.

[4] 段艳杰，吕宜生，张杰，等.深度学习在控制领域的研究现状与展望[J].自动化学报，2016，42(5)：643–654.

[5] 陈硕.深度学习神经网络在语音识别中的应用研究[D].广州：华南理工大学，2013.

[6] 刘大伟，韩玲，韩晓勇.基于深度学习的高分辨率遥感影像分类研究[J].光学学报，2016，36(4)：298–306.

[7] 张建明，詹智财，成科扬，等.深度学习的研究与发展[J].江苏大学学报(自然科学版)，2015，36(2)：191–200.

[8] 孙吉贵，刘杰，赵连宇.聚类算法研究[J].软件学报，2008，19(1)：48–61.

[9] 李晓黎，刘继敏，史忠植.基于支持向量机与无监督聚类相结合的中文网页分类器[J].计算机学报，2001，24(1)：62–68.

[10] 于剑，程乾生.模糊聚类方法中的最佳聚类数的搜索范围[J].中国科学：技术科学，2002，32(2)：274–280.

[11] 李洁，高新波，焦李成.基于特征加权的模糊聚类新算法[J].电子学报，2006，34(1)：89–92.

【责任编辑　梅欣丽】

A New Clustering Analysis Algorithm Based on Deep Learning

HOU Yuanshao

(Department of Information Engineering, Henan Industry and Trade Vocational College, Zhengzhou 451191, China)

Abstract: In the traditional image classification algorithm, a large amount of data was required for supervised training, which caused redundancy of data and brought about the problem of dimensionality disaster. Thus, a clustering analysis algorithm based on deep learning was proposed. Clustering analysis was an unsupervised process. It used the internal cluster structure and pattern of data to classify. It didn't need to train the sample to obtain prior knowledge, which reduced the computational complexity. Deep learning was also introduced to study the internal structure and pattern of the data, and the preliminary clustering of the data was obtained. Then the preliminary clustering was continuously optimized to obtain the final classification effect. Experiments showed that the new algorithm solved the contradiction between comprehensive information and dimensionality disaster, and possessed good practicability and subjective consistency.

Keywords: deep learning; feature extraction; clustering analysis; unsupervised learning