监督学习中的损失函数及应用研究

邓建国，张素兰，张继福，荀亚玲，刘爱琴
太原科技大学计算机科学与技术学院，山西 太原 030024

摘要
监督学习中的损失函数常用来评估样本的真实值和模型预测值之间的不一致程度，一般用于模型的参数估计。受应用场景、数据集和待求解问题等因素的制约，现有监督学习算法使用的损失函数的种类和数量较多，而且每个损失函数都有各自的特征，因此从众多损失函数中选择适合求解问题最优模型的损失函数是相当困难的。研究了监督学习算法中常用损失函数的标准形式、基本思想、优缺点、主要应用以及对应的演化形式，探索了它们适用的应用场景和可能的优化策略。本研究不仅有助于提升模型预测的精确度，而且也为构建新的损失函数或改进现有损失函数的应用研究提供了一个新的思路。

关键词
监督学习；损失函数；相似度度量

中图分类号：TP181 文献标识码：A doi: 10.11959/j.issn.2096-0271.2020006

Loss function and application research in supervised learning

DENG Jianguo, ZHANG Sulan, ZHANG Jifu, XUN Yaling, LIU Aiqin
School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China

Abstract
The loss function in supervised learning is often used to evaluate the degree of inconsistency between the real value of the sample and the predicted value of the model, and is generally used for parameter estimation of the model. Due to the constraints of application scenarios, data sets and problems to be solved, there are many kinds and quantities of loss functions used by existing supervised learning algorithms, and each loss function has its own characteristics. Therefore, it is very difficult to select a loss function suitable for solving the optimal model of the problem from many loss functions. The standard forms, basic ideas, advantages and disadvantages, main applications and corresponding evolution forms of commonly used loss functions in supervised learning algorithms were studied, and their more appropriate application scenarios and possible optimization strategies were summarized. This study not only helps to improve the accuracy of model prediction, it also provides a new idea for the application of new loss functions or to improve the application of existing loss functions.

Key words
supervised learning, loss function, similarity measure

1 引言

随着人工智能在智能制造、智慧农业以及智慧教育等领域的广泛应用，机器学习变得越来越普及，并逐渐成为人工智能研究的重点内容。机器学习以数据为研究内容，使计算机能够自动地从数据中学习规律，并利用规律预测未知数据。机器学习分为监督学习、无监督学习和强化学习。作为机器学习的一个重要类别，监督学习与机器学习同时产生，并伴随着机器学习逐步发展起来。监督学习常用于解决分类或回归问题，是目前研究和应用较为广泛的一种机器学习方法。由于监督学习具有很好的分类和标记能力，被广泛应用于计算机视觉、自然语言、语音识别、目标检测、药物发现和基因组学等多个领域。

监督学习、无监督学习和强化学习都是从数据集中寻找规律，不同的是监督学习从有标记的训练数据集中学习规律，并利用学到的规律预测训练集外的数据的标记，不能预测数据集本身的潜在规律，这在一定程度上限制了监督学习的应用范围。但是，现有的机器学习算法中大多还是基于监督学习的，甚至部分无监督学习和强化学习算法也是基于监督学习并受监督学习思想启发发展起来的。另外，尽管机器学习在多个领域取得了令人瞩目的成绩，但现有机器学习算法的很多结论是通过实验或经验获得的，还有待理论的深入研究与支持。现有机器学习算法无法从根本上解决机器学习面临的技术壁垒，这导致机器学习无法跨越弱人工智能，仍然要依赖监督学习。

监督学习利用有标记的样本调整模型参数，使模型具有正确预测未知数据的能力，其目的是让计算机学习一组有标记的训练数据集，进而获得新的知识或技能，这就要求计算机不断学习样本数据，并依据样本真实值与预测值之间的损失调整模型参数，提升模型的判别能力。显然，衡量样本真实值和预测值不一致程度的损失函数是监督学习研究的重点内容。损失函数是统计学、经济学和机器学习等领域的基础概念，它将随机事件或与其相关的随机变量的取值映射为非负实数，用来表示该随机事件的风险或损失的函数。在监督学习中，损失函数表示单个样本真实值与模型预测值之间的偏差，其值通常用于衡量模型的性能。现有的监督学习算法不仅使用了损失函数，而且求解不同应用场景的算法会使用不同的损失函数。研究表明，即使在相同场景下，不同的损失函数度量同一样本的性能时也存在差异。可见，损失函数的选用是否合理直接决定着监督学习算法预测性能的优劣。

在实际问题中，损失函数的选取会受到许多约束，如机器学习算法的选择、是否有离群点、梯度下降的复杂性、求导的难易程度以及预测值的置信度等。目前，没有一种损失函数能完美处理所有类型的数据。在同等条件下，模型选取的损失函数越能扩大样本的类间距离、减小样本的类内距离，模型预测的精确度就越高。实践表明，在同一模型中，与求解问题数据相匹配的损失函数往往对提升模型的预测能力起着关键作用。因此，如果能正确理解各种损失函数的特性，分析它们适用的应用场景，针对特定问题选取合适的损失函数，就可以进一步提高模型的预测精度。

2 损失函数

监督学习问题是在假设空间F中选取模型f作为决策函数，对于给定的输入x，

用损失函数$L(Y,f(x))$度量该样本经决策函数f计算后的输出预测值$f(x)$与样本真实值Y之间的不一致程度[1]。损失函数是经验风险函数的核心部分，也是结构风险函数的重要组成部分。结构风险最小化策略认为结构风险最小的模型是最优模型，因此求最优模型，就是求解最优化问题：$\min_{f\in F}\frac{1}{N}\sum_{i=1}^{N}L(y_i,f(x_i))+\lambda J(f)$，其中，$J(f)$为模型的复杂度，$\lambda$为系数（$\lambda\geq 0$）。显然，监督学习问题被转化为一个经验风险或结构风险函数的最优化问题[1]。

在监督学习中，损失函数用于评估单个样本经模型计算后输出的预测值与真实值的不一致程度。它是一个非负实值函数，主要特点为：恒非负；误差越小，函数值越小；收敛快。损失函数的值直接影响着模型的预测性能，损失函数值越小，模型的预测性能就越好。另外，作为样本间相似度度量标准，损失函数用来刻画样本真实值与预测值之间的关系，如果损失值小于某一值，则认为样本是相似的，否则认为是不相似的。监督学习算法中的损失函数如图1所示。

损失函数的标准数学形式（以下简称标准公式）不仅种类多，而且每类损失函数又在其标准形式的基础上演化出许多演化形式。0-1损失函数是最简单的损失函数，在其基础上加入参数控制损失范围，形成感知机损失函数；加入安全边界，演化为铰链损失函数。为解决多分类问题，在铰链损失函数的基础上，加入参数k，组合成top-k铰链损失函数。将对数损失函数与softmax函数的特性结合，构成softmax损失函数；与概率分布相似性融合，构成交叉熵（cross entropy）损失函数。另外，组合不同损失函数的标准形式或演化形式又形成新的损失函数。可见，损失函数的发展不是孤立的，而是随着应用研究的发展进行变革的。

本文依据损失函数度量方式的不同，将主要损失函数分为基于距离度量的损失函数和基于概率分布度量的损失函数。同时，进一步研究了每一类损失函数的基本思想、优缺点、演化形式及演化动机，总结了它们的应用场景、更适用的数据集、可能的优化方向，使监督学习的应用研究尽可能选取最优损失函数，以提高模型预测的精确度。同时，给出了监督学习算法中使用频次低的损失函数和组合损失函数。

3 主要损失函数

3.1 基于距离度量的损失函数

基于距离度量的损失函数通常将输入

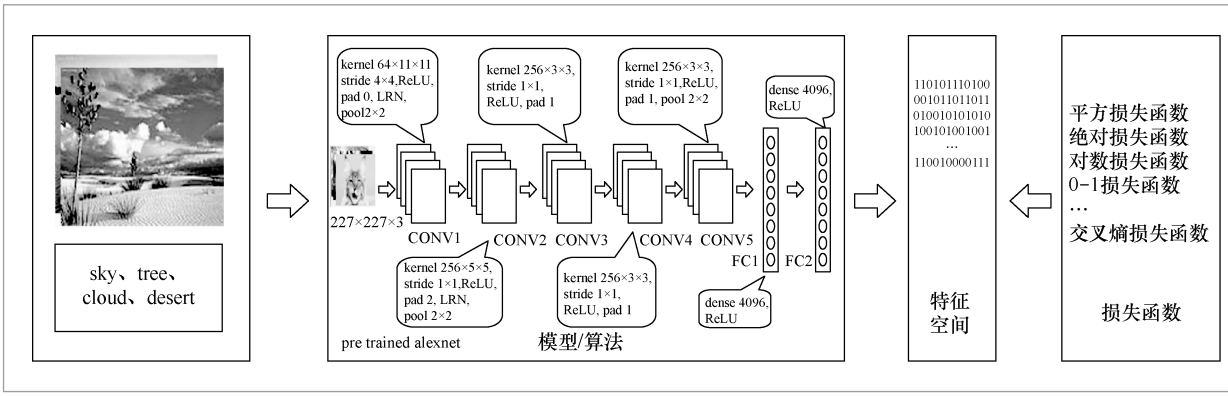

图1 监督学习算法中的损失函数

数据映射到基于距离度量的特征空间上，如欧氏空间、汉明空间等，将映射后的样本看作空间上的点，采用合适的损失函数度量特征空间上样本真实值和模型预测值之间的距离。特征空间上两个点的距离越小，模型的预测性能越好，常用的基于距离度量的损失函数见**表1**。

（1）平方损失函数

平方损失（squared loss）函数最早是从天文学和地理测量学领域发展起来的，后来，由于欧氏距离在各个领域的广泛使用，平方损失函数日益受到研究人员的关注。在回归问题中，平方损失用于度量样本点到回归曲线的距离，通过最小化平方损失使样本点可以更好地拟合回归曲线。在机器学习的经典算法（反向传播算法、循环神经网络、流形学习、随机森林和图神经网络）中，常选用平方损失及其演化形式作为模型对误检测样本进行惩罚的依据。

由于平方损失函数具有计算方便、逻辑清晰、评估误差较准确以及可求得全局最优解等优点，一直受到研究人员的广泛关注，其演化形式也越来越多。基于平方损失演化的损失函数有加权平方损失函数、和方误差（sum squared error，SSE）函数、均方误差（mean squared error，MSE）函数、L2损失（L2 loss）函数、均方根误差（root mean squared error，RMSE）函数、x^2检验（chi-square test）函数、triple损失函数和对比损失（contrastive loss）函数，见**表2**。

加权平方损失函数通过加权修改样本真实值与预测值之间的误差，使样本到拟合曲线的距离尽可能小，即找到最优拟合曲线。在正负样本比例相差很大时，SSE通过计算拟合数据和原始数据对应点的误差平方和，使正样本点更加靠近拟合曲线，与MSE相比，SSE可以更好地表达误差。MSE的思想是使各个训练样本点到最优拟合曲线的距离最小，常用于评价数据的变化程度。MSE的值越小，表示预测模型描述的样本数据具有越好的精确度。由于无参数、计算成本低和具有明确物理意义等优点，MSE已成为一种优秀的距离度量方法。尽管MSE在图像和语音处理方面表现较弱，但它仍是评价信号质量的标准[12]，在回归问题中，MSE常被作为模型的经验损失或算法的性能指标。L2损失又被称为欧氏距离，是一种常用的距离度量方法，通常用于度量数据点之间的相似度。由于L2损

表1 基于距离度量的损失函数

名称	标准形式
平方损失函数	$(Y-f(x))^2$
绝对损失函数	$\lvert Y-f(x)\rvert$
0-1损失函数	$L(Y,f(x))=\begin{cases}1,Y\neq f(x)\\0,Y=f(x)\end{cases}$
铰链损失函数	$\max(0,1-Yf(x)),Y\in\{-1,1\}$
中心损失函数	$\frac{1}{2}\sum_{i=1}^{n}D\left(f(x^i),c_{y_i}\right)$
余弦损失函数	$\frac{Yf(x)}{\lVert Y\rVert\lVert f(x)\rVert}$

表2 基于平方损失的演化损失函数

名称	演化形式
加权平方损失函数	$\lambda(Y-f(x))^2$
和方误差函数	$\sum_{i=1}^{n}(Y_i-f(x_i))^2$
均方误差函数	$\frac{1}{n}\sum_{i=1}^{n}(Y_i-f(x_i))^2$
L2损失函数	$\sqrt{\sum_{i=1}^{n}(Y_i-f(x_i))^2}$
均方根误差函数	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-f(x_i))^2}$
x^2检验函数	$\frac{1}{2}\sum_{i=0}^{n-1}\frac{(A_i-B_i)^2}{A_i+B_i}$
triple损失函数	$\sum_{i}^{N}(\lVert f(x_i^a)-f(x_i^p)\rVert_2^2-\lVert f(x_i^a)-f(x_i^n)\rVert_2^2+\alpha)$
对比损失函数	$\frac{1}{2n}\sum_{i=1}^{N}yd^2+(1-y)\max(\text{margin}-d,0)^2$

失具有凸性和可微性，且在独立、同分布的高斯噪声情况下，它能提供最大似然估计，使得它成为回归问题、模式识别、图像处理中最常使用的损失函数[3]。RMSE直观地揭示了模型的输出预测值与样本真实值的离散程度，常被作为回归算法的性能度量指标。尽管与平均绝对误差（MAE）相比，RMSE计算更复杂且易偏向更高的误差，但由于其是平滑可微的函数，且更容易进行运算，目前仍是许多模型默认的度量标准。x^2检验也被称为x^2统计，常用于计算图像直方图之间的距离。triple损失函数是一个三元损失函数，使用时需设计样本本身、相似的正样本和不相似的负样本三方数据，既耗时又对性能敏感[4]。triple损失函数不能对每一个单独的样本进行约束，由于其具有类间距离大于类内距离的特性，首次出现在基于卷积神经网络（convolutional neural network，CNN）的人脸识别任务中便取得了令人满意的效果[5]。对比损失函数是一个成对损失函数，在使用时除需样本本身外，还需一个对比数据，可见，它也不能对每一个单独的样本进行约束。对比损失函数不仅能降维，而且降维后成对样本的相似性保持不变，可以很好地表达成对样本的匹配程度[6]，另外，它能扩大类间距离，缩小类内距离，在人脸验证算法中，常被作为人脸判断的依据[7]。

总之，平方损失函数不仅常用于回归问题，而且也可用于分类或标注问题，实现离散问题的预测。因为它对离群点比较敏感，所以它不适合离群点较多的数据集。在实际应用中，由于模型泛化问题等原因，一般不常使用平方损失函数的标准形式，而更多使用它的演化形式。另外，由于平方损失函数是可微的凸函数，常与之搭配的优化方法为随机梯度下降或牛顿法。尽管平方损失或基于平方损失的演化损失函数已被广泛应用于图像自动标注、图像重建、对象计数及图像检索等领域，并取得了令人满意的效果，但在选取时也应根据具体问题的实现细节用其优势避其劣势[8]。

（2）绝对损失函数

绝对损失（absolute loss）函数是最常见的一种损失函数，它不仅形式简单，而且能很好地表达真实值和预测值之间的距离。绝对损失对离群点有很好的鲁棒性，但它在残差为零处却不可导。绝对损失的另一个缺点是更新的梯度始终相同，也就是说，即使很小的损失值，梯度也很大，这样不利于模型的收敛。针对它的收敛问题，一般的解决办法是在优化算法中使用变化的学习率，在损失接近最小值时降低学习率。尽管绝对损失本身的缺陷限制了它的应用范围，但基于绝对损失的演化形式却受到了更多的关注。在有噪声标签的分类问题中，基于平均绝对误差构建的神经网络具有良好的噪声容忍能力[9]，在单幅图像超分辨率重建中，选用绝对损失函数可使重建的图像失真更少[10]，smooth L1损失在目标检测问题中可有效解决梯度爆炸问题。

基于绝对损失的演化损失函数包括平均绝对误差函数、平均相对误差（mean relative error，MRE）函数、L1损失（L1 loss）函数、Chebyshev损失函数、Minkowski损失函数、smooth L1损失函数、huber损失函数和分位数损失（quantile loss）函数，见**表3**。

MAE表达预测误差的实际情况，只衡量预测误差的平均模长，不考虑方向，一般作为回归算法的性能指标。MRE既指明误差的大小，又指明其正负方向。一般来说，与MAE相比，它更能反映评估的可信程度。L1损失又称为曼哈顿距离，表示残差的绝对值之和。Chebyshev损失也称切比雪夫距离或L_∞度量，是向量空间中的一种度量方法。Minkowski损失也被称为闵

氏距离或闵可夫斯基距离，是欧氏空间中的一种度量方法，常被看作欧氏距离和曼哈顿距离的一种推广。smooth L1损失是由Girshick R[11]在Fast R-CNN中提出的，主要用在目标检测中防止梯度爆炸。huber损失是平方损失和绝对损失的综合，它克服了平方损失和绝对损失的缺点，不仅使损失函数具有连续的导数，而且利用MSE梯度随误差减小的特性，可取得更精确的最小值。尽管huber损失对异常点具有更好的鲁棒性，但是，它不仅引入了额外的参数，而且选择合适的参数比较困难，这也增加了训练和调试的工作量。分位数损失的思想是通过分位数γ惩罚高估和低估的预测值，使其更接近目标值的区间范围，当设置多个γ值时，将得到多个预测模型，当$\gamma=0.5$时，分位数损失相当于MAE。分位数损失易构建能够预测输出值范围的模型，与MAE相比，它可减少数据预处理的工作量。基于分位数损失的回归学习，不仅适用于正态分布的残差预测问题，而且对于具有变化方差或非正态分布残差的预测问题，也能给出合理的预测区间。在神经网络模型、梯度提升回归器和基于树模型的区间预测问题中，选取分位数损失评估模型的性能往往会取得更好的预测结果。分位数损失函数选取合适的分位数比较困难，一般情况下，分位值的选取取决于求解问题对正误差和反误差的重视程度，应根据实验结果进行反复实验后再选取。另外，分位数损失值在0附近的区间内存在导数不连续的问题。

总之，在实际问题中，一般的数据集或多或少存在离群数据，当离群数据较多或需要考虑离群数据时，利用绝对损失及其演化损失对异常点鲁棒性的特点，可以取得更好的效果。也就是说，绝对损失及其演化损失更适用于有较多离群点的数据集。

表3 基于绝对损失的演化损失函数

名称	演化形式
平均绝对误差函数	$\frac{1}{n}\sum_{i=1}^{n}\lvert Y_i-f(x_i)\rvert$
平均相对误差函数	$\frac{1}{n}\sum_{i=1}^{n}\frac{\lvert Y_i-f(x_i)\rvert}{Y_i}$
L1损失函数	$\sum_{i=1}^{n}\lvert Y_i-f(x_i)\rvert$
Chebyshev损失函数	$\max_{i=1}^{n}(\lvert Y_i-f(x_i)\rvert)$
Minkowski损失函数	$(\sum_{i=1}^{n}\lvert Y_i-f(x_i)\rvert^p)^{\frac{1}{p}}$
smooth L1损失函数	$\begin{cases}\frac{1}{2}(Y-f(x))^2, \lvert Y-f(x)\rvert<1\\ \lvert Y-f(x)\rvert-\frac{1}{2}, \lvert Y-f(x)\rvert\geqslant 1\end{cases}$
huber损失函数	$\begin{cases}\frac{1}{2}(Y-f(x))^2, \lvert Y-f(x)\rvert\leqslant\delta\\ \delta\lvert Y-f(x)\rvert-\frac{1}{2}\delta^2, \lvert Y-f(x)\rvert>\delta\end{cases}$
分位数损失函数	$\sum_{i:y_i<f(x_i)}(1-\gamma)\lvert Y_i-f(x_i)\rvert+\sum_{i:y_i\geqslant f(x_i)}\gamma\lvert Y_i-f(x_i)\rvert$

（3）0-1损失函数

0-1损失（zero-one loss）函数是一种较为简单的损失函数，常用于分类问题。它不考虑预测值和真实值的误差程度，是一种绝对分类方法。其思想是以分隔线为标准，将样本集中的数据严格区分为0或1。由于没有考虑噪声对现实世界数据的影响因素，对每个误分类样本都施以相同的惩罚，预测效果不佳，甚至出现严重误分类情况，这在很大程度上限制了0-1损失函数的应用范围。另外，0-1损失函数是一种不连续、非凸且不可导函数，优化困难，这进一步限制了它的应用范围。基于0-1损失的分类思想，出现了最常见的分类模型——*K*最近邻（*K* nearest neighbor，*K*NN）。虽然0-1损失很少出现在监督学习算法中，但它的分类思想为之后出现的其他分类算法奠定了基础。由于0-1损失函数直观简单，也易理解，研究人员在0-1损失的基础上引入参数，进一步放宽分类标准，将其演化为

感知机损失(perceptron loss)。

感知机损失是0-1损失改进后的结果,它采用参数克服了0-1损失分类的绝对性。与0-1损失相比,它的分类结果更可靠。感知机损失被广泛应用于图像风格化、图像复原等问题中,通过使用预训练的深度网络对图像进行多层语义分解,在相关问题上取得了较好的效果[12],其形式为: $L(Y,f(x))=\begin{cases}1,|Y-f(x)|>t\\0,|Y-f(x)|\leqslant t\end{cases}$,其中$t$为参数,在感知机算法中$t$=0.5。

总之,尽管0-1损失函数存在误分类的情况,但是,当所有样本完全远离分隔线时,0-1损失函数可能是最好的选择,也就是说,0-1损失函数在对称或均匀噪声的数据集上具有较好的鲁棒性。

(4)铰链损失函数

铰链损失(hinge loss)也被称为合页损失,它最初用于求解最大间隔的二分类问题。铰链损失函数是一个分段连续函数,当Y和$f(x)$的符号相同时,预测结果正确;当Y和$f(x)$的符号相反时,铰链损失随着$f(x)$的增大线性增大。铰链损失函数最著名的应用是作为支持向量机(support vector machine,SVM)的目标函数,其性质决定了SVM具有稀疏性,也就是说,分类正确但概率不足1和分类错误的样本被识别为支持向量,用于划分决策边界,其余分类完全正确的样本没有参与模型求解。SVM基本模型是定义在特征空间上间隔最大的线性分类器,当采用核技术后,SVM可转化为非线性分类器。铰链损失函数是一个凸函数,因此,铰链损失函数可应用于机器学习领域的很多凸优化方法中。

基于铰链损失的演化损失函数包括边界铰链损失函数、坡道损失(ramp loss)函数、Crammerand铰链损失函数、Weston铰链损失函数、二分类支持向量机损失函数、多分类支持向量机损失函数、多分类支持向量机平方损失函数和top-k铰链损失函数,见**表4**。

边界铰链损失表示期望正确预测的得分高于错误预测的得分,且高出边界值margin,它主要用于训练两个样本之间的相似关系,而非样本的类别得分。在坡道损失函数中,s为截断点的位置,一般情况下,s的值取决于类别个数c,$s=-\frac{1}{c-1}$,它在x=1和x=s两处不可导。Crammerand铰链损失函数是由Crammerand Singer提出的一种针对线性分类器的损失函数。Weston铰链损失函数是由Weston和Watkins提出的一种损失函数。二分类支持向量机损失函数可看作L2正则化与铰链损失之和。多分类支持向量机损失函数只考虑在正确值附近的那些值,其他的均作为0处理,即只关注那些可能造成影响的点(或支持向量),因此,具有较好的鲁棒性。与多分类支持向量机损失函数相比,多分类支持向量机平方损失函数的惩罚更强烈。top-k铰链损失函数在k个测试样本预测为正的约束下,使所有训练实例的铰链损失最小化[13]。

总之,铰链损失及其演化损失函数常作为人脸识别、文本分类、笔迹识别和图像自动标注领域的损失函数,用于度量图像的相似性或向量空间中向量的距离,对错误越大的样本,它施以越严重的惩罚,这可能使它对噪声敏感,从而降低模型的泛化能力。

(5)中心损失函数

中心损失(center loss)函数采用了欧氏距离思想,为每个类的深层特征学习一个中心(一个与特征维数相同的向量),但在全部样本集上计算类中心相当困难,常用的做法是把整个训练集划分成若干个小

表4 基于铰链损失的演化损失函数

名称	演化形式
边界铰链损失函数	$\max\left(0,\text{margin}-\left(f(x)-Y\right)\right)$
坡道损失函数	$\frac{1}{n}\sum_{i=1}^{n}\left(\max(0,1-h_{y_i})-\max(0,s-h_{y_i})\right)$
crammerand铰链损失函数	$\max(0,1+\max_{Y\neq f(x)} w_Y x-w_{f(x)}x)$
weston铰链损失函数	$\sum_{Y\neq f(x)}\max(0,1+w_Y x-w_{f(x)}x)$
二分类支持向量机损失函数	$\frac{1}{2}\|w\|^2+c\sum_{i=1}^{n}\max(0,1-Yf(x))$
多分类支持向量机损失函数	$\sum_{j\neq Y_i}\max(0,s_j-s_{Y_i}+1)$
多分类支持向量机平方损失函数	$\sum_{j\neq Y_i}\max(0,s_j-s_{Y_i}+1)^2$
top-k铰链损失函数	$C\sum_{i=1}^{n}\left(\max(1-y_i(w^{\mathrm{T}}x_i+b),0)\right)$

的训练集(mini-batch),并在mini-batch样本范围内进行类中心计算。另外,在更新类中心时常增加一个类似学习率的参数α,用于处理采样太少或者有离群点的情况。中心损失函数主要用于减小类内距离,表面上只是减少了类内距离,但实际上间接增大了类间距离,它一般不单独使用,常与softmax损失函数搭配使用,其分类效果比只用softmax损失函数更好。

基于中心损失的演化损失函数有三元中心损失(triplet-center loss,TCL)函数,TCL函数使样本与其对应的中心之间的距离比样本与其最近的负中心之间的距离更近[14],其形式为:$L=\sum_{i=1}^{M}\max\left(D(f_i,c_{y_i})+m-\min_{j\neq y_i}D(f_i,c_j),0\right)$。

总之,中心损失函数常用于神经网络模型中,实现类内相聚、类间分离。在特征学习时,当期望特征不仅可分,而且必须差异大时,通常使用中心损失函数减小类内变化,增加不同类特征的可分离性。在实际应用中,由于中心损失函数本身考虑类内差异,因此中心损失函数应与主要考虑类间的损失函数搭配使用,如softmax损失、交叉熵损失等。

(6)余弦损失

余弦损失(cosine loss)也被称为余弦相似度,它用向量空间中两个向量夹角的余弦值衡量两个样本的差异。与欧氏距离相比,余弦距离对具体数值的绝对值不敏感,而更加注重两个向量在方向上的差异。在监督学习中,余弦相似度常用于计算文本或标签的相似度。它常与词频-逆向文件频率(term frequency-inverse document frequency,TF-IDF)算法结合使用,用于文本挖掘中的文件比较。在数据挖掘领域,余弦损失常用于度量集群内部的凝聚力。

基于余弦损失的演化损失函数有改进的余弦距离核函数,它是由李为等人[15]在与文本相关的说话人确认技术中提出的,用来区分说话人身份及文本内容的差异,其形式为:$L(\lambda(u_1),\lambda(u_2))=\frac{\widetilde{w}(u_1)T\widetilde{w}(u_2)}{\left\|\widetilde{w}(u_1)T\right\|\left\|\widetilde{w}(u_2)\right\|}\times\frac{L(u_1)TL(u_2)}{\left\|L(u_1)T\right\|\left\|L(u_2)\right\|}$,其中,$\lambda$为扬声器模型,$u$为说话人。

3.2 基于概率分布度量的损失函数

基于概率分布度量的损失函数是将样本间的相似性转化为随机事件出现的可能性，即通过度量样本的真实分布与它估计的分布之间的距离，判断两者的相似度，一般用于涉及概率分布或预测类别出现的概率的应用问题中，在分类问题中尤为常用。监督学习算法中，常用的基于概率分布度量的损失函数见**表5**。

（1）对数损失函数

对数损失（logarithm loss）也被称为对数似然损失，它使用极大似然估计的思想，表示样本x在类别y的情形下，使概率$P(y|x)$达到最大值。因为概率的取值范围为$[0,1]$，使得$\log(P(y|x))$取值为$(-\infty,0)$，为保证损失为非负，对数损失的形式为对数的负值。对数损失函数是逻辑回归、神经网络以及一些期望极大估计的模型经常使用的损失函数，它通过惩罚错误的分类，实现对分类器的精确度量化和模型参数估计，对数损失函数常用于度量真实条件概率分布与假定条件概率分布之间的差异。

基于对数损失函数的演化形式包括逻辑回归损失（logistic regression loss）函数、加权对数损失函数、对数双曲余弦损失（log-cosh loss）函数、softmax损失函数、二分类对数损失函数和巴氏距离（Bhattacharyya distance）函数，见**表6**。

逻辑回归损失函数假设样本服从伯努利分布，利用极大似然估计的思想求得极值，它常作为分类问题的损失函数。加权对数损失函数主要应用在类别样本数目差距非常大的分类问题中，如边缘检测问题（边缘像素的重要性比非边缘像素大，可针对性地对样本进行加权）。对数双曲余弦损失函数基本上等价于MSE函数，但又不易受到异常点的影响，是更加平滑的损失函数，它具有huber损失函数的所有优点，且二阶处处可导。softmax损失函数是卷积神经网络处理分类问题时常用的损失函数。巴氏距离函数用于度量两个连续或离散概率分布的相似度，它与衡量两个统计样本或种群之间的重叠量的巴氏系数密切相关。在直方图相似度计算中，选用巴氏距离函数会获得很好的效果，但它的计算很复杂。

总之，在分类学习中，当预测问题使用已知的样本分布，找到最有可能导致这种分布的参数值时，应选取对数损失函数或其演化损失函数作为预测问题的损失函数。在基于深度神经网络的分类或标注问题中，一般在输出层使用softmax作为损失函数，对数损失函数是回归、决策树、深度神经网络常使用的损失函数[16]。

（2）KL散度函数

KL散度（Kullback-Leibler divergence）也被称为相对熵，是一种非对称度量方法，常用于度量两个概率分布之间的距离。KL散度也可以衡量两个随机分布之间的距离，两个随机分布的相似度越高的，它们的KL散度越小，当两个随机分布的差别增大时，它们的KL散度也会增大，因此KL散度可以用于比较文本标签或图像的相似性。基于KL散度的演化损失函数有JS散度函数。JS散度也称JS距离，用于

表5 基于概率分布度量的损失函数

名称	标准形式
对数损失函数	$-\log P(Y\|X)$
KL散度函数	$\sum_{i=1}^{n}P(x_i)\times\log(\frac{P(x_i)}{Q(x_i)})$
交叉熵损失函数	$-\sum_{i=1}^{N}\sum_{j=1}^{C}p_{ij}\log q_{ij}$
softmax损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{f_{y_i}}}{\sum_{j=1}^{c}e^{f_j}}$

表 6　基于对数损失的演化损失函数

名称	演化形式
逻辑回归损失函数	$L(y,P(Y=y\mid x))=\begin{cases}\log(1+\exp(-f(x))), y=1\\ \log(1+\exp(f(x))), y=0\end{cases}$
加权对数损失函数	$-\sum_{k=0}^{C}W_C Y_C \log(f(x))$
对数双曲余弦损失函数	$\sum_{i=1}^{n}\log(\cosh(f(x_i)-Y_i))$
softmax损失函数	$-Y\log(\text{soft}\max(Y,f(x)))$
二分类对数损失函数	$-\frac{1}{n}\sum_{i=1}^{n}(y_i\log p_i+(1-y_i)\log(1-p_i))$
巴氏距离函数	$-\ln\left(BC(p,q)\right)$

衡量两个概率分布之间的相似度，它是基于KL散度的一种变形，消除了KL散度非对称的问题，与KL散度相比，它使得相似度判别更加准确。JS散度函数的形式为：$\text{JS}=\frac{1}{2}\text{KL}\left(P(X)\|\frac{P(X)+Q(X)}{2}\right)+\frac{1}{2}\text{KL}\left(Q(X)\|\frac{P(X)+Q(X)}{2}\right)$，其中，$\text{KL}(\bullet)$为KL散度。

KL散度及其演化损失主要用于衡量两个概率分布之间的相似度，常作为图像低层特征和文本标签相似度的度量标准。KL散度在成像分析[17]、流体动力学[18]、心电图等临床实验室检测、生物应用的网络分析[19]、细胞生物学等领域有广泛的应用。

（3）交叉熵损失

交叉熵是信息论中的一个概念，最初用于估算平均编码长度，引入机器学习后，用于评估当前训练得到的概率分布与真实分布的差异情况。为了使神经网络的每一层输出从线性组合转为非线性逼近，以提高模型的预测精度，在以交叉熵为损失函数的神经网络模型中一般选用tanh、sigmoid、softmax或ReLU作为激活函数。

基于交叉熵损失的演化损失函数包括平均交叉熵损失函数、二分类交叉熵损失函数、二分类平衡交叉熵损失函数、多分类交叉熵损失函数和focal损失函数，见**表7**。

二分类交叉熵损失函数对于正样本而

表 7　基于交叉熵的演化损失函数

名称	演化形式
平均交叉熵损失函数	$-\frac{1}{n}\sum_{i=1}^{n}Y_i\ln(a_i)$
二分类交叉熵损失函数	$-\frac{1}{n}\sum_{i=1}^{n}[Y_i\ln(a_i)+(1-Y_i)\ln(1-a_i)]$
二分类平衡交叉熵损失函数	$-\frac{1}{n}\sum_{i=1}^{n}[\beta Y_i\ln(a_i)+(1-\beta)(1-Y_i)\ln(1-a_i)]$
多分类交叉熵函数	$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}Y_{i,j}\log(a_{i,j})$
focal损失函数	$-\frac{1}{n}\sum_{i=1}^{n}Y_i\partial_i(1-a_i)^{\gamma}\log(a_i)$

言，输出概率越大，损失越小；对于负样本而言，输出概率越小，损失越小。二分类平衡交叉熵损失函数与二分类交叉熵损失函数相比，它的优势在于引入了平衡参数$\beta \in [0,1]$，可实现正负样本均衡，使预测值更接近于真实值。

交叉熵损失函数刻画了实际输出概率与期望输出概率之间的相似度，也就是交叉熵的值越小，两个概率分布就越接近，特别是在正负样本不均衡的分类问题中，常用交叉熵作为损失函数。目前，交叉熵损失函数是卷积神经网络中最常使用的分类损失函数，它可以有效避免梯度消散。

（4）softmax损失函数

从标准形式上看，softmax损失函数应归到对数损失的范畴，在监督学习中，由于它被广泛使用，所以单独形成一个类别。softmax损失函数本质上是逻辑回归模型在多分类任务上的一种延伸[20]，常作为CNN模型的损失函数。softmax损失函数的本质是将一个k维的任意实数向量$\boldsymbol{x}$映射成另一个k维的实数向量，其中，输出向量中的每个元素的取值范围都是(0,1)，即softmax损失函数输出每个类别的预测概率。由于softmax损失函数具有类间可分性，被广泛用于分类、分割、人脸识别、图像自动标注和人脸验证等问题中，其特点是类间距离的优化效果非常好，但类内距离的优化效果比较差。

基于softmax损失函数的演化损失函数包括softer softmax函数、NSL（normalized softmax loss）函数、LMCL（large margin cosine loss）函数、L-softmax函数、A-softmax函数、AM-softmax（additive margin softmax）函数以及正则化softmax函数，见**表8**。

表 8 基于 softmax 损失的演化损失函数

名称	演化形式
softer softmax损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\frac{e^{\frac{f_{Y_i}}{T}}}{\sum_{j=1}^{c}e^{\frac{f_j}{T}}}$
NSL损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s\cos(\theta_{Y_i},i)}}{\sum_{j\neq Y_i}e^{s\cos(\theta_j,i)}}$
LMCL损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s(\cos(\theta_{Y_i},i)-m)}}{e^{s(\cos(\theta_{Y_i},i)-m)}+\sum_{j=1,j\neq Y_i}e^{s\cos(\theta_j,i)}}$
L-softmax损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{\|W_{Y_i}\|\|x_i\|\psi(\theta_{Y_i})}}{e^{\|W_{Y_i}\|\|x_i\|\psi\cos(\theta_{Y_i})}+\sum_{j\neq Y_i}e^{\|W_j\|\|x_i\|\cos(\theta_j)}}$
A-softmax损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{\|x_i\|\psi(\theta_{y_i},i)}}{e^{\|x_i\|\psi(\theta_{y_i},i)}+\sum_{j=1,j\neq Y}^{c}e^{\|x_i\|\cos(\theta_j,i)}}$
AM-softmax损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s(W_{Y_i}^{T}f_i-m)}}{e^{s(W_{Y_i}^{T}f_i-m)}+\sum_{j=1,j\neq Y_i}^{c}e^{sW_j^{T}f_i}}$
正则化softmax损失函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{f_{Y_i}}}{\sum_{j=1}^{c}e^{f_j}}+\lambda R(W)$

softer softmax 损失函数是Hinton G等人[21]为了解决模型给误分类标签分配的概率被softmax损失忽略的问题而提出的。NSL损失函数利用两个特征向量之间的余弦相似度评估两个样本之间的相似性，使后验概率只依赖于角度的余弦值，由此产生的模型学习了角空间中可分离的特征，提高了特征学习能力。为了构建一个更大边距的分类器，期望$\cos(\theta_1)-m>\cos(\theta_2)$且$\cos(\theta_2)-m>\cos(\theta_1)$，$m\geqslant 0$为控制余弦边矩大小的参数，Wang H等人[22]提出了LMCL损失函数。与传统的欧几里得边距相比，角的余弦值与softmax具有内在的一致性，基于此思想，在原始softmax基础上选用角边距度量两个样本的相似性，Liu W等人[4]提出了L-softmax损失函数。受L-softmax损失函数启发，在它的基础上，Liu W等人[23]增加了条件$\|W\|=1$，$B=0$和$\cos(m\theta_1)>\cos(\theta_2)$，使得预测仅取决于$W$和$x$之间的角度$\theta$，提出了A-softmax损失函数。受L-softmax损失函数的启发，Wang F等人[24]提出了AM-softmax损失函数，它使前后向传播变得更加简单。正则化softmax损失函数加入刻画模型的复杂度指标的正则化，可以有效地避免过拟合问题。

softmax损失函数具有类间可分性，在多分类和图像标注问题中，常用它解决特征分离问题。在基于卷积神经网络的分类问题中，一般使用softmax损失函数作为损失函数，但是softmax损失函数学习到的特征不具有足够的区分性，因此它常与对比损失或中心损失组合使用，以增强区分能力[23]。

4 其他损失函数

其他损失函数主要包括指数损失（exponential loss）函数、汉明距离函数、dice损失函数、余弦损失+softmax函数和softmax+LDloss函数，见**表9**。

与主要损失函数相比，其他损失函数在监督学习中使用的频次比较低，其中指数损失函数是AdaBoost算法中常用的损失函数。它与铰链损失函数和交叉熵损失函数相比，对错误分类施加的惩罚更大，这使得它的误差梯度也较大，因此在使用梯度下降算法优化时，在极小值处求解速度也较快。汉明距离用于计算两个向量的相似度，即通过比较两个向量的每一位是否相同来计算汉明距离。dice损失函数是一种集合相似性度量函数，通常用于计算两个样本的相似性，常作为文本比较或图像分割类问题的损失函数，尤其适用于处理图像的前景区域和背景区域相差较大的图像分割问题。余弦损失+softmax函数是利用余弦和softmax的特性组合而成的，见第3.2节中的LMCL损失函数。softmax+LDloss函数是黄旭等人[25]在融合判别式深度特征学习的图像识别算法中引入线性判别分析（linear discriminant analysis，LDA）思想构建的损失函数，该算法使softmax+LDloss参与卷积神经网络的训练，实现尽可能最小化类内特征距离和

表9 其他损失函数

名称	标准形式
指数损失函数	$\exp(-Yf(x))$
汉明距离函数	$\sum_{i=1}^{n} I(x_i,y_i)$
dice损失函数	$1-\frac{\sum_{i=1}^{n} p_i r_i+\varepsilon}{\sum_{i=1}^{n} p_i+r_i+\varepsilon}-\frac{\sum_{i=1}^{n}(1-p_i)(1-r_i)+\varepsilon}{\sum_{i=1}^{n} 2-p_i-r_i+\varepsilon}$
余弦损失+softmax函数	$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s(\cos(\theta_{Y_i},i)-m)}}{e^{s(\cos(\theta_{Y_i},i)-m)}+\sum_{j=1,j\neq Y_i} e^{s\cos(\theta_j,i)}}$
softmax+LDloss函数	$L_s+\alpha L_L$

最大化类间特征距离的目标，以提高特征的鉴别能力，进而改善图像的识别性能。

5 损失函数在监督学习中的应用

在监督学习中，损失函数的选取和演化通常是伴随着监督学习算法的发展而演变的。监督学习算法是通过计算机学习有标记的训练数据集，使模型能对未知数据进行预测的机器学习任务。在机器学习发展的过程中，产生了许多监督学习算法，其中，典型算法有深度神经网络（deep neural network，DNN）、决策树（decision tree，DT）、朴素贝叶斯分类、线性回归（linear regression）、逻辑回归、支持向量机、*K*最近邻和AdaBoost等，这些监督学习算法仍是当前研究和应用的重点内容。

5.1 深度神经网络中的应用

深度神经网络是深度学习的一个重要分支，由于深度神经网络具有更多的网络层数和参数个数，所以能够提取更多的数据特征，获取更好的学习效果。在基于深度神经网络的应用中，损失函数作为度量样本真实值和模型预测值之间差异的工具，主要解决视觉处理、自然语言和语音识别等领域的问题。在视觉识别应用中，中心损失函数可区分人脸的差异[26]、提高人脸识别准确度[27]；softmax损失函数结合fisher损失函数提升不同模态下人脸数据的关联程度[28]；改进的triplet损失函数解决了跨摄像机人员再识别问题[29]。在图像标注应用中，利用JS散度生成独特的图像标签[30]，采用交叉熵损失函数构建分层标签的图像分类模型[31]，选用余弦损失函数表达语义相似性[32]。在视觉检索应用中，通过L2损失函数构建用于大规模视觉搜索的深度Hash方法[33]。在图像重建方面，采用均方误差函数重建出准确度更高的图像[34]。在图像分割应用中，在softmax分类器后增加dice损失函数，以提高图像分割精度[35]。在语音识别应用中，使用A-softmax损失函数提高端到端系统的性能[36]。在自然语言处理应用中，使用加权交叉熵生成富含情感的单词[37]。

5.2 决策树中的应用

决策树是一个预测模型，一般依据KL散度大小，将样本属性分解成树状结构，并用于新样本分类。决策树的3种典型实现为ID3、CART和C4.5。决策树的计算复杂度低，可解释性强，对中间值的缺失不敏感，这使得决策树至今在一些问题上仍被使用。在基于决策树的应用中，引入C4.5决策树方法处理流量分类问题，利用训练数据集中的KL散度构建分类模型，并通过对分类模型的简单查找实现未知网络流样本的分类，理论分析和实验结果表明，使用C4.5决策树处理流量分类问题在分类稳定性方面具有明显的优势[38]。在视觉位置识别问题中，将汉明距离嵌入二叉搜索树中，解决描述符匹配和图像检索问题[39]。

5.3 朴素贝叶斯分类中的应用

朴素贝叶斯分类算法是一个典型的机器学习算法，它假定样本的各个特征之间是相互独立的，在大量样本下会有较好的表现，不适用于与输入向量的特征条件有关联的场景。由于朴素贝叶斯分类算法实现简单，并有坚实的数学理论作为支撑，因此在很多领域有广泛的应用，如垃圾邮件过滤、文本分类等。在基于朴素贝叶斯分类的

应用中，采用余弦损失度量成对标签相似度提高标签在整个图像相似性分配上的一致性[40]，汉明损失函数和0-1损失函数在多标签分类中分类能力不同[41]，0-1损失函数在有限样本下具有良好的分类效果[42]。

5.4 线性回归中的应用

线性回归是监督学习中经典的回归模型之一，它是利用数理统计中的回归分析，确定两个或两个以上变量间相互依赖的定量关系的一种统计分析方法。线性回归分为一元线性回归和多元线性回归。因为线性回归形式简单、易于建模，常用于解决连续值预测的问题。在基于线性回归的应用中，引入KL散度估计簇间距离，不仅能够获得更精确的分割结果，而且对噪声和初始轮廓具有更强的鲁棒性[43]；在存在异常值或重尾误差分布的情况下，基于指数平方损失函数的线性回归估计方法比最小二乘数估计方法更有效[44]；采用均方误差函数可精确估计交通拥挤时的车辆数量[45]；自适应huber损失函数不仅解决了原始huber损失函数没有封闭解、难以优化的问题，而且在曲线拟合、带噪声标签图像标注、经典回归问题和人群计数应用方面有很大的优势[46]。

5.5 逻辑回归中的应用

逻辑回归是一个应用非常广泛的机器学习算法，它将数据拟合到一个逻辑函数中，预测事件发生的概率。逻辑回归使用极大似然估计思想，常选用对数损失或交叉熵损失作为损失函数，可用于解决二分类或多分类问题。在基于逻辑回归的应用中，采用基于对数损失的核函数度量两个样本之间的相似性设计的逻辑回归模型，与核逻辑回归分析和支持向量机相比，不仅可达到更好的分类精度，而且有更好的时间效率[47]；在研究分类不平衡对软件缺陷影响的问题中，采用对数损失函数的逻辑回归模型的性能更加稳定[48]；采用最小化对数损失构建的点击通过率（click-through rate，CTR）预测模型，比传统的点击率预测模型以及最新的基于深度学习的预测模型的性能更好[49]。

5.6 支持向量机中的应用

支持向量机基于最大化分类间隔的原则，通过核函数巧妙地将线性不可分问题转化为线性可分问题，并且具有非常好的泛化性能。它使用铰链损失函数计算经验风险，并在求解系统中加入了正则化项以优化结构风险，是一个具有稀疏性和稳健性的分类器，它在文本分类和图像标注领域有广泛的应用。在基于支持向量机的应用中，采用JS散度计算两个特征向量之间的相似度，在原始SVM基础上，引入概率加权策略构建多个分类器，设计了基于SVM的多特征融合的图像标注方法[50]；采用余弦损失函数计算向量空间模型（vector space model）中向量的相似度，提出了内容和标签相融合的图像标注方法[51]；在标签混淆情况下，将公差参数引入铰链损失函数中，可提高标签分类准确率[52]。

5.7 *K*最近邻算法中的应用

*K*最近邻（*K*NN）算法是简单的机器学习算法之一，该算法的思想是如果一个样本在特征空间中的*K*个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别样本的特性。*K*NN算法在进行类别决策时，只与极少量的相邻样本有关，因此，它适用于类

域有交叉或重叠较多的待分类样本集。在基于*K*NN算法的应用中，采用铰链损失函数度量两幅图像之间的相似性进行归纳学习，采用余弦损失函数度量两幅图像标签的相似性进行推理学习，并将归纳学习和推理学习结合，形成统一的距离度量学习框架[53]；在经典的2PKNN算法中，采用铰链损失函数度量两幅图像的相似性，解决了类别不平衡和弱标记问题[54]；利用汉明距离度量*K*NN搜索上的可伸缩性，实现了跨模态相似度搜索[55]。

5.8 AdaBoost算法中的应用

AdaBoost算法是基于Boosting思想的集成学习算法，使用指数函数作为损失函数，其核心思想是对同一个训练集训练不同的弱学习器，然后将多个弱学习器进行集成，构造一个精度非常高的强学习器。AdaBoost算法被广泛应用于计算机视觉和目标检测领域，在人脸检测方面的表现尤为出众。在基于AdaBoost算法的应用中，通过学习图像局部区域的相似性得到一组非线性弱学习器，使用弱学习器训练图像局部特征获得低维的、独有的特征描述子，选用汉明距离度量特征描述子之间的相似度，提高了图像局部区域的匹配精度[56]。利用改进的指数损失函数与平方损失函数的加权组合代替传统AdaBoost指数损失函数构建模型，解决传统的AdaBoost算法在训练样本存在异常值时导致的分类效果不理想的问题[57]。

6 结束语

本文依据损失函数对样本的评估方式，将监督学习的主要损失函数分为基于距离度量的损失函数和基于概率分布度量的损失函数，并分析了每个损失函数的基本思想、优缺点、主要应用，在此基础上，总结了每个损失函数更适合的应用场景或可能的优化方向。另外，在其他损失函数中给出了在监督学习算法中出现但使用频次相对较低的损失函数和组合函数。在监督学习中，损失函数作为度量数据真实值与预测值之间相似度的工具，它直接影响着模型的预测性能，显然，损失函数的选取或改进是监督学习领域研究的主要内容。

本文提到的主要损失函数在监督学习算法中使用频次比较高，对于具体的研究问题，这些损失函数不一定是最好的损失函数，可能存在其他更合适的损失函数。其原因是模型的产生与解决的具体问题有关，同一模型在同类问题中性能差异不大，但在相似类问题或不同类问题中，性能差异可能较大，甚至出现不适合的情形。导致此类问题产生的原因除模型本身问题外，另一个主要原因是模型默认的损失函数。尽管每个模型都有默认的损失函数，但在使用模型解决实际问题时，应考虑问题与损失函数的内在联系，尤其是使用当下流行的迁移学习解决问题时，更应考虑问题与损失函数的关系。

虽然监督学习已产生多年，但在机器学习中仍占有重要地位，其主要原因是监督学习模型在解决问题方面的精确度一直在不断提高，甚至在某些方面的能力已经超越人类。模型精确度提升的原因除模型结构等因素不断改进外，另一主要原因是不断优化的损失函数在度量预测值与真实值差异方面的能力在不断提升，甚至在不改变模型结构的情形下，只调整损失函数就可以达到很好的效果。相比于模型结构的升级换代，损失函数的优化工作量要少

得多，可见优化损失函数也是提升模型性能的有效手段。优化损失函数除在现有标准形式的基础上扩展延伸外，还应考虑组合不同损失函数形成组合损失函数。组合损失函数是将现有的损失函数经过四则运算，构造一类新的损失函数，笔者认为组合损失函数构成的一般原则是基于问题性质，以损失函数最小为目标，确定预测模型中各单项预测的加权系数，并将各个单独的损失函数组合成一个新的损失函数，利用新的损失函数度量样本之间的差异。

构建监督学习算法的目的是解决实际问题，而实际问题中的主要内容是数据。在监督学习模型中，损失函数通过度量数据来衡量模型的性能，可见损失函数与数据直接相关。不同类的问题中数据间差异较大，甚至是不相关的，而相似类问题中的数据差异往往不大，甚至存在共同特性，这为实际问题选定的模型选择合适的损失函数提供了参考。笔者认为应首先考虑问题中数据的性质，然后找到解决此类问题效果好的模型，在选定模型的基础上，进一步考虑损失函数与数据的内在关系，进而选择更合适的损失函数。

尽管损失函数在模型中的地位比较重要，但它毕竟只是模型的一个部分，要使模型有更好的预测性能，在实际问题中还需要考虑其他影响模型性能的因素，如数据特征归一化方法、样本间相似度度量方法、相近类别分离技术以及组合损失函数等。总之，监督学习是一个系统工程，除研究损失函数外，还应考虑数据集、模型结构及优化策略等，只有围绕损失函数全面考虑影响它的因素，才会取得更好的预测效果。

通常情况下，损失函数的选取应从以下方面考虑：

- 选择最能表达数据的主要特征构建基于距离或基于概率分布度量的特征空间；
- 选择合理的特征归一化方法，使特征向量转换后仍能保持原来数据的核心内容；
- 选取合理的损失函数，在实验的基础上，依据损失不断调整模型的参数，使其尽可能实现类别区分；
- 合理组合不同的损失函数，发挥每个损失函数的优点，使它们能更好地度量样本间的相似性；
- 将数据的主要特征嵌入损失函数，提升基于特定任务的模型预测精确度。

下一步可能的工作有两方面：一是在现阶段研究的基础上，进一步研究监督学习算法中的损失函数，以期能找到更普适的度量样本特征的损失函数；二是研究无监督学习和强化学习中的损失函数，总结其中典型的损失函数，并尝试将其引入监督学习应用中，以提高模型预测的精确度。

本文在常用损失函数的基础上，为构建新的损失函数或改进现有损失函数的应用研究提供了一个新的思路，不仅有助于研究人员针对研究问题轻松选择适合问题的损失函数，而且还可根据待求解问题改进或优化现有的损失函数。总之，监督学习是当前的一个热门研究领域，也是未来很有前途的一个研究方向，对监督学习算法中的损失函数研究具有重大的理论意义和应用前景，是一个具有实用价值的研究课题。

参考文献：

[1] 李航. 统计学习方法[M]. 北京：清华大学出版社，2012.
LI H. Statistical learning method[M].

2020006-16

Beijing: Tsinghua University Press, 2012.

[2] ZHOU W, BOVIK A C. Mean squared error: love it or leave it? A new look at signal fidelity measures[J]. IEEE Signal Processing Magazine, 2009, 26(1): 98-117.

[3] ZHAO H, GALLO O, FROSIO I, et al. Loss functions for image restoration with neural networks[J]. IEEE Transactions on Computational Imaging, 2017, 3(1): 47-57.

[4] LIU W, WEN Y, YU Z, et al. Large-margin softmax loss for convolutional neural networks[C]//The 33rd International Conference on Machine Learning, June 19-24, 2016, New York, USA. [S.l.:s.n.], 2016: 507-516.

[5] SCHROFF F, KALENICHENKO D, PHILBIN J. Facenet: a unified embedding for face recognition and clustering[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. Piscataway: IEEE Press, 2015: 815-823.

[6] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 17-22, 2006, New York, USA. Piscataway: IEEE Press, 2006: 1735-1742.

[7] SUN Y, CHEN Y, WANG X, et al. Deep learning face representation by joint identification-verification[C]//The 27th International Conference on Neural Information Processing Systems, December 13, 2014, Montreal, Canada. New York: ACM Press, 2014: 1988-1996.

[8] 刘思琦, 郎丛妍, 冯松鹤. 基于对抗式扩张卷积的多尺度人群密度估计[J]. 中国图象图形学报, 2019, 24(3): 483-492.
LIU S Q, LANG C Y, FENG S H. Multi-scale crowd counting via adversarial dilated convolutions[J]. Journal of Image and Graphics, 2019, 24(3): 483-492.

[9] GHOSH A, KUMAR H, SASTRY P S. Robust loss functions under label noise for deep neural networks[C]//The 31st AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, USA. Palo Alto: AAAI Press, 2017: 1919-1925.

[10] 应自炉, 龙祥. 多尺度密集残差网络的单幅图像超分辨率重建[J]. 中国图象图形学报, 2019, 24(3): 410-419.
YING Z L, LONG X. Single-image super-resolution construction based on multi-scale dense residual network[J]. Journal of Image and Graphics, 2019, 24(3): 410-419.

[11] GIRSHICK R. Fast R-CNN[C]//The IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. Piscataway: IEEE Press, 2015: 1440-1448.

[12] 李国庆, 赵洋, 刘青萌, 等. 多层感知分解的全参考图像质量评估[J]. 中国图象图形学报, 2019, 24(1): 153-162.
LI G Q, ZHAO Y, LIU Q M, et al. Multi-layer perceptual decomposition based full reference image quality assessment[J]. Journal of Image and Graphics, 2019, 24(1): 153-162.

[13] LIU L P, DIETTERICH T G, LI N, et al. Transductive optimization of top k precision[C]//The 25th International Joint Conference on Artificial Intelligence, July 9-15, 2016, New York, USA. Palo Alto: AAAI Press, 2016: 1781-1787.

[14] HE X, ZHOU Y, ZHOU Z, et al. Triplet-center loss for multi-view 3D object retrieval[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, USA. Piscataway: IEEE Press, 2018: 1945-1954.

[15] 李为, 游寒旭, 朱杰, 等. 一种应用于文本相关说话人确认的L-向量表示和改进的余弦距离核函数[J]. 上海师范大学学报:自然科学版, 2016(2): 243-247.
LI W, YOU H X, ZHU J, et al. A novel

L-vector representation and improved cosine distance kernel for text-dependent speaker verification[J]. Journal of Shanghai Normal University(Natural Sciences), 2016(2): 243-247.

[16] PAINSKY A, WORNELL G. On the universality of the logistic loss function[C]//2018 IEEE International Symposium on Information Theory, June 17-22, 2018, Vail, USA. Piscataway: IEEE Press, 2018: 936-940.

[17] LIANG S, YANG F, WEN T, et al. Nonlocal total variation based on symmetric Kullback-Leibler divergence for the ultrasound image despeckling[J]. BMC Medical Imaging, 2017, 17(1): 57.

[18] GRANERO-BELINCHÓN C, ROUX S G, GARNIER N B. Kullback-Leibler divergence measure of intermittency: application to turbulence[J]. Physical Review E, 2018, 97(1): 013107.

[19] JERO S E, RAMU P, RAMAKRISHNAN S. Discrete wavelet transform and singular value decomposition based ECG steganography for secured patient information transmission[J]. Journal of Medical Systems, 2014, 10(38): 1-11.

[20] KING G, ZENG L. Logistic regression in rare events data[J]. Political Analysis, 2001, 9(2): 137-163.

[21] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. Computer Science, 2015, arXiv:1503.02531.

[22] WANG H, WANG Y, ZHOU Z, et al. Cosface: large margin cosine loss for deep face recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, USA. Piscataway: IEEE Press, 2018: 5265-5274.

[23] LIU W, WEN Y, YU Z, et al. SphereFace: deep hypersphere embedding for face recognition[C]//The IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, USA. Piscataway: IEEE Press, 2017: 212-220.

[24] WANG F, CHENG J, LIU W, et al. Additive margin softmax for face verification[J]. IEEE Signal Processing Letters, 2018, 25(7): 926-930.

[25] 黄旭, 凌志刚, 李绣心. 融合判别式深度特征学习的图像识别算法[J]. 中国图象图形学报, 2018, 23(4): 510-518.
HUANG X, LING Z G, LI X X. Discriminative deep feature learning method by fusing linear discriminant analysis for image recognition[J]. Journal of Image and Graphics, 2018, 23(4): 510-518.

[26] WEN Y, ZHANG K, LI Z, et al. A discriminative feature learning approach for deep face recognition[C]//European Conference on Computer Vision. Springer, October 8-16, 2016, Amsterdam, The Netherlands. Heidelberg: Springer, 2016: 499-515.

[27] WEN Y, ZHANG K, LI Z, et al. A comprehensive study on center loss for deep face recognition[J]. International Journal of Computer Vision, 2019, 127(6-7): 668-683.

[28] 董震, 裴明涛. 基于异构哈希网络的跨模态人脸检索方法[J]. 计算机学报, 2019, 42(1): 75-86.
DONG Z, PEI M T. Cross-modality face retrieval based on heterogeneous Hashing network[J]. Chinese Journal of Computers, 2019, 42(1): 75-86.

[29] CHENG D, GONG Y, ZHOU S, et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 1335-1344.

[30] WU B, CHEN W, SUN P, et al. Tagging

like humans: diverse and distinct image annotation[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, USA. Piscataway: IEEE Press, 2018: 7967-7975.

[31] HU H, ZHOU G T, DENG Z, et al. Learning structured inference neural networks with label relations[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 2960-2968.

[32] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013, arXiv:1301.3781.

[33] ERIN LIONG V, LU J, WANG G, et al. Deep hashing for compact binary codes learning[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, USA. Piscataway: IEEE Press, 2015: 2475-2483.

[34] KIM J, KWON LEE J, MU LEE K. Accurate image super-resolution using very deep convolutional networks[C]//The IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, USA. Piscataway: IEEE Press, 2016: 1646-1654.

[35] 詹曙, 梁植程, 谢栋栋. 前列腺磁共振图像分割的反卷积神经网络方法[J]. 中国图象图形学报, 2017, 22(4): 516-522.
ZHAN S, LIANG Z C, XIE D D. Deconvolutional neural network for prostate MRI segmentation[J]. Journal of Image and Graphics, 2017, 22(4): 516-522.

[36] LI Y, GAO F, OU Z, et al. Angular softmax loss for end-to-end speaker verification[C]//2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), November 26-29, 2018, Taipei, China. Piscataway: IEEE Press, 2018: 190-194.

[37] ZHONG P, WANG D, MIAO C. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss[C]//The AAAI Conference on Artificial Intelligence, January 27-February 1, Honolulu, USA. Palo Alto: AAAI Press, 2019: 7492-7500.

[38] 徐鹏, 林森. 基于C4.5决策树的流量分类方法[J]. 软件学报, 2009, 20(10): 2692-2704.
XU P, LIN S. Internet traffic classification using C4.5 decision tree[J]. Journal of Software, 2009, 20(10): 2692-2704.

[39] SCHLEGEL D, GRISETTI G. HBST: a Hamming distance embedding binary search tree for feature-based visual place recognition[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3741-3748.

[40] WANG H, HU J. Multi-label image annotation via maximum consistency[C]//2010 IEEE International Conference on Image Processing, September 26-29, 2010, Hong Kong, China. Piscataway: IEEE Press, 2010: 2337-2340.

[41] DEMBCZYŃSKI K, WAEGEMAN W, CHENG W, et al. Regret analysis for performance metrics in multi-label classification: the case of hamming and subset zero-one loss[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases, September 20-24, 2010, Barcelona, Spain. Heidelberg: Springer, 2010: 280-295.

[42] SHALEV-SHWARTZ S, SHAMIR O, SRIDHARAN K. Learning kernel-based halfspaces with the zero-one loss[J]. Computer Science, 2010, arXiv:1005.3681.

[43] CHENG D , TIAN F , LIU L , et al. Image segmentation based on multi-region multi-scale local binary fitting and

Kullback-Leibler divergence[J]. Signal Image & Video Processing, 2018(2): 1-9.

[44] YU P, ZHU Z, ZHANG Z. Robust exponential squared loss-based estimation in semi-functional linear regression models[J]. Computational Statistics, 2019, 34(2): 503-525.

[45] ONORO-RUBIO D, LÓPEZ-SASTRE R J. Towards perspecti-ve-free object counting with deep learning[C]//European Conference on Computer Vision, October 8-16, 2016, Amsterdam, The Netherlands. Heidelberg: Springer, 2016: 615-629.

[46] CAVAZZA J, MURINO V. Active regression with adaptive huber loss[J]. Computer Science, 2016, arXiv:1606.01568.

[47] 毛毅, 陈稳霖, 郭宝龙, 等. 基于密度估计的逻辑回归模型[J]. 自动化学报, 2014, 40(1): 62-72.
MAO Y, CHEN W L, GUO B L, et al. A novel logistic regression model based on density estimation[J]. Acta Automatica Sinica, 2014, 40(1): 62-72.

[48] 于巧, 姜淑娟, 张艳梅, 等. 分类不平衡对软件缺陷预测模型性能的影响研究[J]. 计算机学报, 2018, 41(4): 809-824.
YU Q, JIANG S J, ZHANG Y M, et al. The impact study of class imbalance on the performance of software defect prediction models[J]. Chinese Journal of Computers, 2018, 41(4): 809-824.

[49] 刘梦娟, 曾贵川, 岳威, 等. 基于融合结构的在线广告点击率预测模型[J]. 计算机学报, 2019, 42(7): 1570-1587.
LIU M J, ZENG G C, YUE W, et al. A hybrid network based CTR prediction model for online advertising[J]. Chinese Journal of Computers, 2019, 42(7): 1570-1587.

[50] WU W, NIE J, GAO G. An improved SVM-based multiple features fusion method for image annotation[J]. Journal of Information & Computational Science, 2014, 11(14): 4987-4997.

[51] CHAN S B, YAMANA H, LE D D, et al. Image annotation fusing content-based and tag-based technique using support vector machine and vector space model[C]//2014 10th International Conference on Signal-Image Technology and Internet-Based Systems, November 23-27, 2014, Marrakech, Morocco. Piscataway: IEEE Press, 2014: 272-276.

[52] VERMA Y, JAWAHAR C V. Exploring SVM for image annotation in presence of confusing labels[C]//British Machine Vision Conference, September 9-13, 2013, Bristol, UK. [S.l.:s.n.], 2013: 1-11.

[53] WU P, HOI S C H, ZHAO P, et al. Mining social images with distance metric learning for automated image tagging[C]// The 4th ACM International Conference on Web Search and Data Mining, February 9-12, 2011, Hong Kong, China. New York: ACM Press, 2011: 197-206.

[54] VERMA Y, JAWAHAR C V. Image annotation using metric learning in semantic neighbourhoods[C]// European Conference on Computer Vision, October 7-13, 2012, Firenze, Italy. Heidelberg: Springer, 2012: 836-849.

[55] ZHAI D, LIU X, CHANG H, et al. Parametric local multiview hamming distance metric learning[J]. Pattern Recognition, 2018, 75(C): 250-262.

[56] 惠国保, 童一飞, 李东波. 基于改进的图像局部区域相似度学习架构的图像特征匹配技术研究[J]. 计算机学报, 2015, 38(6): 1148-1161.
HUI G B, TONG Y F, LI D B. Image features matching based on improved patch similarity learning framework[J]. Chinese Journal of Computers, 2015, 38(6): 1148-1161.

[57] XING H J, LIU W T. Robust AdaBoost based ensemble of one-class support vector machines[J]. Information Fusion, 2020, 55: 45-58.

作者简介

邓建国(1977-),男,太原科技大学计算机科学与技术学院硕士生,主要研究方向为数据挖掘与图像理解。

张素兰(1971-),女,博士,太原科技大学计算机科学与技术学院教授,中国计算机学会(CCF)会员,主要研究方向为粒计算、数据挖掘与图像理解。

张继福(1963-),男,太原科技大学计算机科学与技术学院教授、博士生导师,CCF高级会员,主要研究方向为数据挖掘与高性能计算。

荀亚玲(1980-),女,博士,太原科技大学计算机科学与技术学院副教授,主要研究方向为数据挖掘与并行计算。

刘爱琴(1975-),女,太原科技大学计算机科学与技术学院副教授,主要研究方向为数据挖掘、并行与分布式计算。

收稿日期:2019-11-25

通信作者:张素兰, zhsulan@126.com

基金项目:国家自然科学资金资助项目(No.61373099, No.61602335)

Foundation Items: The National Natural Science Foundation of China (No.61373099, No.61602335)