文章编号:1007-5321(2018)01-0001-12 **DOI**:10.13190/j.jbupt.2017-150

机器学习中的特征选择方法研究及展望

崔鸿雁[1,2,3], 徐 帅[1,2,3], 张利锋[1,2,3], Roy E. Welsch[4], Berthold K. P. Horn[5]

(1. 北京邮电大学 网络与交换技术国家重点实验室, 北京 100876; 2. 北京邮电大学 网络体系构建与融合北京市重点实验室, 北京 100876; 3. 先进信息网络北京实验室, 北京 100876; 4. Sloan School of Management, Massachusetts Institute of Technology, MA 02139, USA; 5. Csail Laboratory, Massachusetts Institute of Technology, MA 02139, USA)

摘要: 任何领域的大数据研究都离不开用机器学习方法提取特征. 为了探求满足海量大数据分析需求的特征选择方法,笔者对利用机器学习进行特征选择的常用方法做了深入分析,归纳总结出特征选择的五大类方法:相关性度量方法、Lasso 稀疏选择方法、集成方法、神经网络方法、主成分分析方法. 通过对比不同特征选择方法的原理、实现过程以及应用场景,给出了不同算法下进行特征选择时的适用范围、优缺点和关键点,为研究者提供参考.

关 键 词: 机器学习; 特征选择; 迁移学习; 对抗神经网络; 人工智能

中图分类号: TN929.53 **文献标志码**: A

The Key Techniques and Future Vision of Feature Selection in Machine Learning

CUI Hong-yan[1,2,3], XU Shuai[1,2,3], ZHANG Li-feng[1,2,3], Roy E. Welsch[4], Berthold K. P. Horn[5]

(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China; 3. Beijing Laboratory of Advanced Information Networks, Beijing 100876, China; 4. Sloan School of Management, Massachusetts Institute of Technology, MA 02139, USA; 5. Csail Laboratory, Massachusetts Institute of Technology, MA 02139, USA)

Abstract: Big data research is widely spread around the world, and feature selection of machine learning plays an important role on these researches. To address the issue of discovering novel feature selection methods in data mining tasks on big data, this paper researches five models related to feature selection: linear coefficient correlation, Lasso sparse selection, ensemble learning models, neural networks, principal component analysis. The merits and drawbacks of these models are extensively discussed in depth in this paper, which may help in providing a direction for those who are interested in the machine learning area.

Key words: machine learning; feature selection; transfer learning; generative adversarial networks; artificial intelligence

随着互联网、大数据、物联网的发展,机器学习 已经被广泛使用在人类生存和生活的各个领域,成

收稿日期: 2017-07-20
基金项目: 教育部-中国移动科研基金项目(MCM20170306)
作者简介: 崔鸿雁(1977—), 女, 博士生导师, E-mail:cuihy@bupt.edu.cn.

为人工智能不可或缺的一部分. 比如 AlphaGo[1] 通过模拟蒙特卡洛等算法,在围棋领域完胜人类选手,掀起了机器学习领域的高潮;自然语言处理中通过机器学习解决输入法的智能联想[2]、机器翻译[3-4]、语音输入[5-6]、文本语义分析和情感分析[7-8];图像识别和图像处理中,用来解决图像分类[9-11]、人脸识别[12-13]、运动物体检测[14-15];协同过滤等算法用于推荐系统,解决活跃用户检测[16]、电商商品推荐[17]、精准广告推荐等[18].

然而,截至目前,大部分机器学习的应用往往受限于单一领域,一个很重要的原因就在于特征的相关性[19-20]. 一个性能优良的模型,依赖于相关度大的特征集合. 在业界流行一个说法:"特征工程决定了泛化能力的上限,模型和算法只不过是不断去趋近这个泛化上限而已". 由此可见,特征工程在整个机器学习过程中相当重要. 大数据在很多情况下是电子形式存储的非结构化数据,哪怕是结构良好的表单数据,表面上的维度十分有限,直接使用这些特征进行机器学习是十分困难的. 第1个难点在于如何从这些维度中提取出更多和需求目标相关的特征,这属于特征抽取和特征挖掘的范畴. 第2个难点在于如何去掉那些相关度不大的特征,将更少的特征应用于机器学习流程,该过程称为特征选择.

特征工程是机器学习的基石,而特征选择是特征工程的一个重要方面. 特征选择主要致力于解决3个问题.

1) 维度灾难[21]与数据需求问题. 在特征抽取的过程中,总是倾向于去生成更多的特征,但是并不是特征越多越好. 越多的特征意味着越大的数据需求,而特征选择可以以指数形式减小数据需求.

2) 过拟合[22]问题. 目前的机器学习任务大多是 NP 难问题,大多采用多项式的模型和算法来完成这些任务. 因此,更多的特征意味着更大的模型复杂度. 根据 Ockham's Razor 原则[23],模型复杂度的增大,会增加模型过拟合的风险.

3)噪声问题. 如果把机器学习解释成一个信号去噪,或者说是信号解码过程,更多的特征给了噪声更多的引入机会,而无关的冗余特征将把信号淹没在噪声的海洋中. 去掉这些冗余特征,既减小了引入噪声的风险,又减小了机器学习迭代过程中的计算量,还可大大减小分布式框架的网络传输开销.

特征选择的方法五花八门. 不同的任务不同的人来做,八仙过海,各显神通,目前没有一个统一的标准,也没有一个系统的步骤,或完整的实现方案,原因在于具体问题的复杂性和机器学习的发展快速. 南京大学的周志华[24]教授在他的《机器学习》一书中提到了3种特征选择方法:嵌入式、过滤式和包裹式. 嵌入式方法是将特征选择的过程作为一部分,融入模型的训练过程中;过滤式方法是在模型训练之前先一个一个地过滤出特征,之后再将选出的这些特征输入模型;而包裹式则是直接使用不同的特征子集来训练不同的结果,根据最优的模型结果来选择最优的特征子集. 显然,嵌入式方法的选择相对方便,但是需要精心设计模型,且输入到模型的数据庞大,计算开销大;过滤式方法逐个选择特征,对于模型的迭代开销较小,但是需要假设特征相互独立,这样就遗漏了特征与特征之间的相关关系;而包裹式方法虽然全面、无遗漏,但是对每个特征子集都需要进行一遍模型训练,对于大数据模型来说可行性较差. 然而,机器学习中的特征选择并不仅仅局限于这3种做法. 一个鲁棒性好的机器学习模型可以是单独的一种做法,也可以是多种特征选择方法的混合,或通过嵌套、堆叠等方式取长补短,亦或通过模型与数据处理交替迭代的方式选择特征.

笔者分析了近30年机器学习中特征选择的相关论文,总结归纳出五大类主流算法:相关性度量方法、Lasso 稀疏选择方法、集成方法、神经网络方法和主成分分析方法. 笔者的主要贡献在于通过深入分析主流特征选择的算法原理、实现过程以及应用场景,归纳总结了各个算法在特征选择方面的优、缺点和适用场景,为人们根据机器学习任务选择合适的特征提取方法提供参考,给研究新的特征选择方法提供思路.

1 用相关性度量进行特征选择

对于一个监督学习任务而言,假设各个特征之间相互独立,某个特征的数值分布与监督值的分布相关度越大,说明这个特征对训练目标的贡献越大,而相关度小的特征可以认为是冗余特征,通过简单的过滤算法就可以进行特征选择. 通过调整相关度的阈值,或者对各个特征排序后取前面一部分可以挑选出那些相关度大的特征,之后再将挑选出来的相关度大的特征输送到机器学习迭代模块进行训练. 这样做有三大好处:1)算法实现简单,对于每个特征只需要计算一次相关度,有时候可以通过采样的方式减少计算量,而特征选择也只需要设计布尔

形式的过滤函数;2)输入到迭代模块的数据量小,大大减少了迭代过程中的数据传输开销和计算开销;3)算法的可解释性和实用性较强. 基于相关性度量的缺点也很明显,它假设特征相互独立,在实际的应用中很难满足这个假设条件,是一种贪婪方法,很难实现全局最优.

在机器学习领域,皮尔逊系数常被用于衡量 2 个分布之间的线性相关性,如式(1)所示,其中分子是 2 个分布的协方差,而分母是 2 个数据分布各自的标准差相乘的积. 从公式中可以看出,皮尔逊系数是对称的,即 $p(X,Y)$ 等于 $p(Y,X)$. 另外,皮尔逊系数的最小值是 -1,最大值是 1. 绝对值越大,表明 2 个分布的线性相关性越大,可以选择那些和监督值相关性大的特征进入模型训练. 皮尔逊系数已被广泛应用于机器学习场景,Jordan 等[25]在高维基因微阵列中基于皮尔逊系数进行特征选择,Singh 等[26]在文本分类任务中使用了皮尔逊系数. 除了皮尔逊系数外,还有其他衡量线性关系的系数,比如余弦相似度、Jaccard 距离、欧氏距离,但是因为皮尔逊系数衡量能力比较全面,适用能力强,应用场景会比其他线性相关系数多一些. 余弦相似度常用于自然语言处理中,比如 2 个文档的相似性比较;Jaccard 距离常见于集合间的相似性比较,比如 2 个路径点集合比较;欧式距离往往与地理空间有关,比如 GPS 定位预测.

$$p(X,Y)=\frac{\mathrm{Cov}(X,Y)}{\sigma(X)\sigma(Y)} \tag{1}$$

虽然大部分线性相关系数计算简单,但是只能反映 2 个数据分布之间的线性关系,而数据之间也存在大量非线性关系. 互信息可以衡量 2 个分布之间的信息量关系,如式(2)所示,其中 p 表示联合概率或者边缘概率. 互信息是对称的且非负的,其物理意义在于一个分布从另一个分布获得的平均信息量,值越大表示相关性越大. 互信息也经常被使用在机器学习领域. Yang 等[27]在文本分类工作中的特征选择利用了互信息, Peng 等[28]将互信息系数用于分析最大依赖、最大相关、最小冗余等工作. 虽然互信息相对于那些线性相关系数能够反映出非线性相关关系,但是互信息的计算量相对较低,对于一些计算能力有限的场景,应该优先考虑线性相关系数. 另外,对于物理意义明确的场景,应该优先考虑相对应的相似性度量.

$$I(X,Y)=\sum_{y\in Y}\sum_{x\in X}p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

进行简单的复杂度分析. 皮尔逊系数、余弦相似度、Jaccard 距离、欧式距离、互信息都可以通过直接计算或者牺牲空间的方式得到线性复杂度. 假设样本容量为 N,有 F 个特征,那么如果不考虑特征组合而逐个筛选特征,需要 $O(NF)$ 的复杂度. 若考虑特征组合,包含空集有 2^F 次方个组合方案,复杂度为 $O(2^FN)$.

无论是线性相关系数,还是非线性相关系数,都有其自身的局限性,只能描述特定方面的相关性. 具体的机器学习任务应该仔细分析任务场景,通过做大量尝试来获得一个较优的度量方式,必要的时候需要自己设计合适的相似性度量,但是需要经过足够数量的验证. 另外有一点需要注意的是,相似性度量应该满足一致性、归一性和单调性. 一致性是指同一个 x 在任何时刻与同一个 y 进行比较,得出的相似性度量应该是一致的. 归一性是指各个参与相似性度量成分的贡献应该处于同一个量级和量纲,比如一个数据样本有 2 个维度,第 1 个维度的取值范围是 0 ~ 1,而第 2 个维度的取值范围是 100 ~ 10 000,如果直接在这类数据上应用欧式距离,那么第 2 个维度将起主要作用,这时就需要先仔细分析第 2 个维度数据分布的情况之后将第二个维度进行某种方式的归一化操作之后再应用欧式距离[29]. 单调性是指物理意义上 z 点离 x 点比 y 点离 x 点要远,那么设计的相似性度量对 z 与 x 比 y 与 x 要小.

2　使用 Lasso 进行稀疏选择

在数学上,少量的数据是 1,大量的数据为 0,称为稀疏数据,类似的概念有稀疏向量、稀疏矩阵、稀疏编码等. 稀疏性的好处在于可以使用高效的存储方式(哈希)来表示高维空间,而稀疏性的坏处在于包含了大量冗余信息. Lasso 方法[30]通过对机器学习模型的目标函数添加一个带权重的惩罚项来选择特征,这个惩罚性约束了模型各个特征上的权重. 通过增大这个惩罚项的权重,使得模型的大量系数趋于 0,从而满足了稀疏性 ,即 L1 正则化. 从表面上看,因为大量特征的系数变成了 0,模型的复杂度大大减小了,从而大大减小了过拟合的风险. 这实质上是在模型训练过程中进行了特征选择,那些系数被训练为 0 的特征被认为是冗余特征,因为这些特征的贡献度是 0. 在模型训练完之后,可以手工去掉这些系数为 0 的冗余特征,将模型上线之后,需要输入的特征数量便大大减小了,后续的数据开销便会

比较小,因为一些数据的采集和传输工作被精简了.可以说 Lasso 是一次训练就完成了特征选择,本身的实现也比较简单.

以线性回归任务为例,Lasso 回归的目标式为

$$\text{L1}: \quad T = \text{Loss}(x,\omega) + \lambda \| w \| \tag{3}$$

$$\text{L2}: \quad T = \text{Loss}(x,\omega) + \lambda \| w \|^2 \tag{4}$$

其中:Loss 为模型的损失函数,一般采用最小二乘损失,也可以是其他损失函数;Lasso 的做法就是在目标式的后面追加一个各特征权重的第一范式作为惩罚项. 只要这个惩罚项前面的系数足够大,在最小化目标式时,惩罚项也得到了优化,使得各个特征的权重趋于0,实现了特征权重的稀疏性. 在实际应用中,可以根据实际的模型性能调整惩罚项前面的系数,调整的方式可以采用网格搜索等方式. 式(4)是更早被提出的岭回归(Ridge)算法[31]的目标式,它与 Lasso 的目标式之间的差别在于惩罚项不同. Lasso 的惩罚项是第一范式 L1,即各个特征权重的绝对值之和,而 Ridge 的惩罚项是第二范式 L2,即特征权重的平方之和. 虽然 L1 和 L2 都能从一定程度上减小特征权重,从而降低模型复杂度,进而降低机器学习模型的过拟合风险,但是 L1 比 L2 范式有更大的稀疏性,而 L2 具有更大的平滑性[32]. L1 范式之所以有更大的稀疏性,直观的解释是 L1 和 L2 的优化等值线形状不一样,如图1所示. 考虑较简单的情况,只有在2个维度上进行优化,第一范式所构造的优化等值线是正方形的形状,而第二范式所构造的等值线是圆形的. 由于正方形比圆形更容易与凸优化方向相交于坐标轴上,而交点就是凸优化的最优值,这样就使在最优情况下某个维度上的值取得了0值,相当于把这个特征过滤掉了,从而造成了特征稀疏,也可以推广到高维情况,而第二范式因交点落在坐标轴上比较难,而具有平滑性.

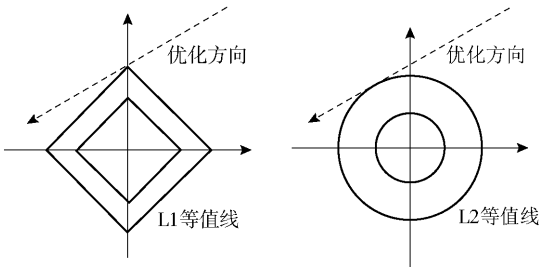

图1 第一范式和第二范式稀疏性比较

Lasso 的目标优化需要使用启发式方法. 因为 Lasso 的惩罚项不是凸的,使整个目标式的优化不是凸的,不能直接借助梯度来解决,但是可以使用近似梯度等方法. 经过算法优化可以让 Lasso 的收敛速度接近于最小二乘法的收敛速度,理论上复杂度最小可以达到 $O(NF^2)$,其中 N 为训练样本数量,F 为特征数量.

使用第一范式和第二范式进行正则化各有优点,ElasticNet 就是将两类惩罚项同时携带权重追加到线性回归模型的目标式中,在优化目标式的过程中对稀疏性和平滑性进行折中[33]. 对于 Lasso 本身的改进算法层出不穷,比较著名的有 Zhou 等[34]提出的 Exclusive Lasso 算法. 基于核技巧,通过让不同任务的相同特征进行竞争,实现多任务特征选择;Efron 等[35]提出的最小角回归算法通过计算残差和相关系数不断调整步长进行前向 Lasso 回归,用于特征选择减小了贪婪算法的缺陷. 由于 L1 范式具有稀疏性、实现简单、可解释性强等诸多优点,常被应用于其他机器学习模型,如极限梯度提升(XGBoost,extreme gradient boosting)模型,相对于传统的梯度提升决策树(GBDT,gradient boosting decision tree)模型,在目标式中增加了 L1 和 L2 等惩罚项,更能减低过拟合风险;在深度神经网络中[36],对每一层网络可以追加 L1 范式,从而让一部分神经元的系数为0,实现网络的稀疏,降低过拟合的风险. 从某种角度来看,神经网络的每一层都是基于上一层的特征提取出更加抽象的一层特征,所以神经网络可以利用 L1 范式进行特征选择.

3 使用集成方法进行特征选择

集成方法主要有两大类:一类是基于 bagging 思想的模型;另一类是基于 boosting 思想的模型. 根据强大数定理,集成学习模型和普通模型不同,随着基础模型的数量增多也不易过拟合. Bagging 方法的出发点和目的是降低模型的方差来提升模型效果,通过 Bootstrap 的方式抽取数据用于独立训练多个基础模型,并将这些基础模型的预测结果通过平均投票的方式给出最终结果. 因为基础模型是独立训练的,bagging 基础的模型可以方便地实现并行计算. 随机森林[37-38]应用了 bagging 基础模型. Boosting 的出发点和目的则是降低偏差来提升模型的效果,先训练一个弱学习器,之后再基于上一轮模型的误差来训练下一个弱分类器,并且累加到整体的模型中. 其中自适应增强(Adaboost,adaptive boosting)算法[39]是通过计算上一轮模型对每个训练数据样

本点的误差，增大那些被误分类数据的权重输入到这一轮训练中，而误差小的弱学习器在累加过程中的权重相应比较大. 而传统的梯度增强回归树(GBRT，gradient boosted regression trees) 算法[40]则每次基于上一轮模型的预测残差来训练新的弱分类器，用新的基础学习器去拟合残差，最后将这些基础学习器组合起来成为一个整体. 传统的 GBRT 是不支持并行计算的，但是近几年的一些 GBRT 已经实现了并行计算，如微软的 Light-GBM 框架，多次在 Kaggle 大赛上拿冠军的 XGBoost 系统[41]. 集成方法既适用于回归任务，也适用于分类任务. 特征选择是嵌入在集成方法训练过程中的，而集成模型中进行特征选择不同于其他特征选择方法的一个重要方面是集成模型进行特征选择是面向最终的模型性能的，正好与特征选择的目的相同.

随机森林是一个很成功的基于 bagging 算法集成模型. 一般采用分类和回归树(CART，classification and regression tree)[42]作为基础模型. 之所以使用 CART 树，是因为 CART 决策树采用 Gini 增益作为树分裂依据，而 C4.5 和 ID3 等决策树采用香农熵增益，而 Gini 增益是对香农熵的有效近似，但计算量却能大大降低. 对于每个基础模型，输入的数据是原始数据经过 Bootstrap 采样的结果. 所谓 Bootstrap 采样，就是从原训练数据集可重复地采样，以构造新的训练数据集，这种采样在理论上约有 1/3 的数据没有被采样出来，这些数据称为袋外数据，可以用于特征重要性的估计. 另外，每棵树并不是采用全部特征，而是随机使用一定量的特征. 一般情况下，如果有 n 个特征，可以采样$\sqrt{n}$个特征，这种技术称为列采样. Bootstrap 技术和列采样技术增加了模型的随机性，大大减低了过拟合风险. 因此，不必对 CART 树进行剪枝，就可减小剪枝开销. 随机森林模型之所以能被广泛地用于特征选择，是因为在完成模型训练之后可以给出各个特征的重要性评价. 对特征的重要性排序，可以筛选出那些相对重要性较高的特征，并且将这些特征应用于再次训练或者其他模型. 具体地说，每个特征的重要性为各个基础学习器上的特征重要性的平均，对于每个基础学习器，先对袋外数据进行预测，计算出一个误差率. 每次对一个特征引入随机白噪声干扰，再利用袋外数据计算误差率，用这 2 个误差率之差反映特征重要性，误差率越大，说明这个特征对于该学习器来说越重要[43]. 随机森林模型在多个领域都有应用，Díaz-Uriart 等[44]利用随机森林进行基因选择和分类，并且获得了几乎无冗余的基因组合. Pal 等[45]将随机森林模型应用于远程感知发现，参数比支持向量机少，而且预测效果和支持向量机持平. Chen 等[46]设计了一种将随机森林模型用于预测蛋白质间的链接，召回率达到了 79.68%.

XGBoost 是基于 GBRT 算法的高效实现，它既是一个机器学习算法，也是一个高效的机器学习系统，并可提供多种编程语言的应用编程接口(API，application programming interface). 目前 Kaggle 大赛上一半以上的冠军都或多或少使用了 XGBoost. XGBoost 也屡屡出现在国内的天池算法赛中. 这个算法相对于传统的 GBRT 有诸多优点，主要有 4 个方面：1) XGBoost 同时使用一、二阶梯度对残差进行近似，其目标式为

$$T^{(t)} \approx \sum_{i=1}^{n}\left[\mathrm{Loss}(y_i,\hat{y}_i^{t-1}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) + \mathrm{constant} \tag{5}$$

其中：y 为拟合值，g 为一阶梯度，h 为二阶梯度，f 为响应函数值，为模型的结构复杂度，constant 代表模型的其他常量. 此时，XGBoost 比仅适用一阶梯度的传统 GBRT 收敛速度快得多；2) 传统的 GBDT 很少考虑正则，而 XGBoost 引入了 L1 和 L2 正则项，降低了模型复杂度和过拟合的风险；3) XGBoost 也引入了列采样技术和学习步长技术，大大降低了过拟合风险；4) XGBoost 是在特征的角度而不是基础学习器的角度支持了并行，同时采用直方图算法加速了树的分裂，对内存需求更友好，在模型训练过程中，自动进行特征选择. XGBoost 一般也是使用 CART 回归树作为基础学习器，但是不限于 CART 树. 与随机森林不同，XGBoost 通过随机白噪声对 Gini 增益的扰动来衡量特征的重要性，Gini 增益变化越大说明特征越重要，应该优先选择那些重要性大的特征. 在 XGBoost 的应用中，它也常常被用于特征选择，如顾客行为预测[47]和商品推荐[48].

简单分析一下集成模型的时间复杂度. 若采用的基础模型是决策树，则需要 $O(FN\log N)$ 复杂度，那么通过 M 棵树做集成，则需要 $O(MFN\log N)$ 的复杂度.

无论是随机森林模型还是 XGBoost，都是在模型训练中无人为干预地进行了特征选择. 此外，集成学习方法也特别适合大规模数据的迭代. 随机森林模型的训练对缺失值不敏感，而 XGBoost 也对带

有缺失值的特征分裂进行了优化.

4 神经网络自动选择特征

2006年,Hinton等[49]在Science上发表了“利用神经网络对数据降维”一文,掀起了神经网络在机器学习领域的研究热潮. 随着深度学习[50-51]的推进,图像处理领域著名的基于卷积神经网络(CNN, convolution neural network)的各种算法得到了快速发展[52-54],文本序列处理领域著名的循环神经网络(RNN,recurrent neural network)、长短期记忆网络(LSTM,long short-term memory)以及他们的各种变种也空前繁荣[55-57]. 理论上,一个隐藏层以上的神经网络只要神经元数量足够多就可以表示任意形式的函数映射关系,而深度网络是指神经网络的深度越来越深,微软公司在2015年的ImageNet大赛上夺冠的残差网络其深度竟达到了152层[58]. 神经网络之所以往深度发展而不往广度发展的一个重要原因是特征的模块化和层次化. 举个最简单的例子,如果要在同一个层次表示8个字符需要8 bit,而层次化之后只需3 bit. 深度神经网络通过设置神经元的层级和链接将特征进行层层抽象,在训练模型的过程中没有人为干预就进行特征抽取和特征选择. 虽然这些特征的可解释性和可读性比较差,但是只要应用效果好,就可以认为选择出了有价值的特征,而大量竞赛的优异结果和实际上线的项目的巨大成功也证明了通过深度神经网络训练方法提取出特征的有效性. 但是深度的加大也带来了许多问题,其中最主要2个问题. 1)数据量的问题. 每加深一个层次,虽然数据特征进行了更深层次的抽象,但是也意味着需要成倍地增加数据才能覆盖这些特征;2)梯度弥散问题. 因为目前大多数深度神经网络采用的是基于链式求导法则的反向传播算法,在后一层向前一层传送梯度增量时,每一层都乘上了一个系数,这个系数很可能小于1. 因此层数多了之后,前面的层次增量几乎为0,在进行梯度优化时前面几层的系数就调不动. 梯度弥散问题一般可以通过逐层训练方法或者Wake-Sleep算法等得到缓解[59],而数据量问题往往需要通过设计巧妙的激活函数、链接方式等减少模型的复杂度和训练难度,从而降低数据需求.

下面以卷积神经网络中最经典的LeNet5为例,介绍用神经网络进行特征提取和特征选择的过程. 图2描述了经典LeNet5的模型架构. 模型的设计目标是解决手写数字识别问题,共有0~9十类手写数字,是一个多分类任务. 每张都是32×32个像素的平面图像,可以对应到一个矩阵,通过一个5×5的卷积核,得到的图像大小是28×28,在第1个卷积层上使用6个卷积核,便能得到6张28×28的图像,将这些图像按顺序堆叠在一起,便得到了28乘以28×6的混合图层. 通过卷积作用将相邻位置加权聚合,当卷积核里的系数都是1时,是滑动求和;当卷积核里的系数都是卷积核大小分之一时就是滑动平均. 一个卷积核就代表了一种聚合特征. 一维卷积很容易推广到二维、三维的卷积作用. 卷积层之后紧接着一个2×2的池化层,如果是maxpooling池化层,就是将每4个值一组采样出其中最大的一个值,这样神经元的数量就减小了1/4,这是因为正则化作用减低了网络层之间的耦合,其原理是基于图像的伸缩位移不变. 之后再接一层卷积层和池化层,接着是3层全连接层,最后使用Softmax函数输出分类标签. 在模型框架搭建好后,通过反向传播算法训练出各个卷积核的系数.

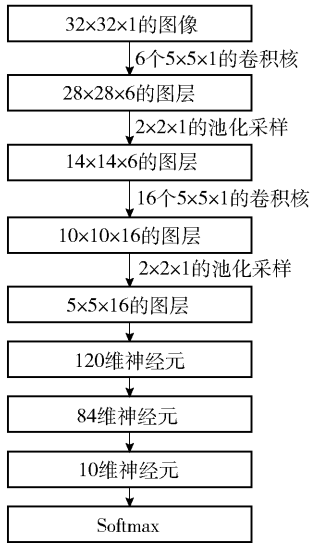

图2 LeNet5神经网络架构

通过分析可以发现,卷积层相对于传统全连接的神经网络,在卷积核上共享大量权重,需要训练的参数数量成级数下降,使得训练任务的难度大大减低,同时需要的数据量也减少了. 可以看出,一个卷积核就代表了一类特征,而每类特征的选择是在模型训练过程中无人为干预的情况下进行的,可见每类特征都为最终的模型效果服务. 需要注意的是,初始化卷积核需要随机取值,而合适的初始化值也将在一定程度上加速训练过程.

在实际应用中，神经网络的训练需要很多技巧. 1）使用线性整流函数（ReLU，rectified linear unit）激活单元[60]能够大大简化计算量，减小梯度弥散，制造稀疏性，但是 ReLU 单元容易死掉；2）使用 L1 和 L2 范式对每 1 层进行正则化，减小过拟合；3）使用 batch-normalization[61]，将数据按批进行归一化，可以缓解梯度问题；4）使用 dropout 技术[62]，按一定比例随机删除一些神经元，可以造成稀疏性和增大随机性，降低过拟合风险. 而深度卷积神经网络架构上的改进也非常多：例如 2012 年的 ILSVRC 比赛第 1 名的 AlexNet 一共设计了 8 层架构，采用了 2 个图形处理单元（GPU，graphics processing unit）进行计算，并且将网络拼接在一起，前 5 个分类的错误率达到了 15.3%[63]；2014 年 ImageNet 大赛上第 1 名 GoogleNet 一共有 22 层结构，所设计网络中的子结构实现了局部稀疏，前 5 个分类的错误率下降到 6.67%[64]；2015 年在 ImageNet 大赛上的冠军 ResNet 残差网络使用了 152 层，跨层连接使得网络实现了更深的层次，将前 5 个分类的错误率降至 3.57%[65].

虽然现在只需要分配充足的硬件资源和设计合理的网络框架就可以以很低的错误率处理图像分类任务，但是仍然需要大量的数据来支持训练. 如果数据量不够大怎么办？2014 年 Goodfellow 等[66]提出了一种生成对抗网络，可以在一定程度上解决数据量的问题. 因为生成对抗网络可以自动地从已有数据中提取和选择出数据特征，并且利用这些数据特征伪装出相类似的数据样本. 如图 3 所示.

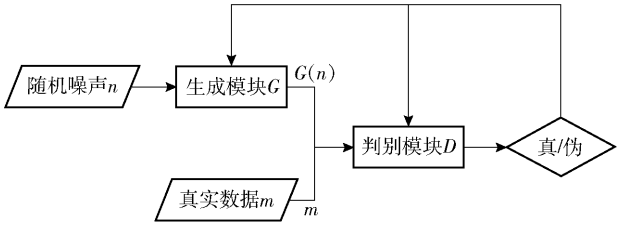

图 3　GAN 网络架构

生成对抗网络架构由生成模块和判别模块组成. 判别模块用于判断输入的图像是否足够真实. 初始模块均未经训练，两组模块不断训练，生成模块连续生成新的结果，并用判别模块来检测，两者的能力均互相促进，最终达到一个生成样本和原始样本无限趋近的状态. 显然，这是 2 个模块相互对抗的过程，对于生成模块，尽量去伪装真实数据，从而通过组合这些特征来生成伪造的数据. 而判别模块则是尽量去判别出能真实描述数据原貌的特征，以区分哪些是真实的数据，哪些是伪造的数据. 虽然生成对抗网络在框架上是很清晰的，但是在实现上有比较多的难点. 如要考虑马太效应，不能让模型一边倒，生成模块效果一直很差，而判断模块的效果一直很好. 这就需要一个能动态平衡的目标函数. 当模型进行目标优化时，判别模块和生成模块可以保持一种动态平衡. 一种可行的目标函数见（6）

$$T^{D}(\theta_D,\theta_G)=-\frac{1}{2}E_m[\log D(m)]-\frac{1}{2}E_n[\log(1-D(G(n)))] \tag{6}$$

其中：m 为真实数据，n 为生成的伪造数据，G 为生成模块，D 表示判别模块，E 表示期望. 通过对目标函数进行优化，见式（7），可以进行模型对抗生成训练.

$$\min_G\max_D\{f(D,G)=E_m[\log D(m)]+E_n[\log(1-D(G(n)))]\} \tag{7}$$

截至目前，对抗生成网络已经被广泛用于机器学习的各个领域，它能够根据原始图像生成无限接近真实数据所构成的图像. Twitter 公司的 Ledig 等[67]利用视觉几何小组（VGG，visual geometry group）的 VGGNet网络作为判别模块，残差网络 ResNet 作为生成模块，通过高斯滤波器，并采用下采样的方式提取高分辨率图像，对于每个图像，作者随机裁剪 16 个 96×96 像素的高分辨率子图作为训练图像，成功地将一个低像素模糊图像还原为高像素清晰图像，图像特征被放大了 4 倍. 对抗生成神经网络还可以用于从文本制造图像. 在 Reed 等[68]的研究中生成模型的输入为文本数据，生成器与判别器均利用深度卷积网络提取文本特征，将文本特征嵌入 1 024 维的图像，交替更新生成器和判决器，然后用产生的文本特征生成图像. 判决器通过描述文本的嵌入特征，在深度上与图片进行拼接，卷积操作后，计算得分来判断相关度. 对抗生成网络也可以用于文本的语义理解. Monroe 等[69]在对话机器人训练过程中使用对抗神经网络技术，该生成模型使用了 seq2seq 的判别模型，判别模型使用了分层编码，只在正序列 y + 以及负序列 y - 的各个子序列里面随机地提取一个样本来训练判别器 D，从而实现生成对话.

将对抗生成网络用于特征选择具有众多优点，主要包括以下几个方面：1）网络框架适用于任意一类的生成器学习网络；2）不必创设依照各类因式分

解的模型,训练过程为2个神经网络的不断对抗,训练过程进行特征提取,增大了精准度,缩短了启动时间;3)可以使用反向传播算法进行训练迭代,整个过程无需用马尔科夫链去反复采样,无需推测学习,回避了计算概率复杂的问题,提升了生成模型的训练效率. 但是,对抗生成神经网络也有一些缺陷亟待解决,主要包括以下3个方面:1)在凸函数的斜率下降时,可以确保达成纳什均衡;在实际未达成均衡时,若两边博弈均由人工神经网络呈现,可以实现让模型持续保持调整策略模式;2)虽然对抗生成网络生成的数据各种各样,但是存在模式瓦解的问题,有时不满足收敛性;3)对抗生成网络开始时不用建模,模型随意发展容易失去控制,不一定能拟合真实数据.

下面简单分析一下神经网络的时间复杂度. 假设有 N 个训练样本,L 个链接,每个链接前向激活的时间开销是 a,后向反馈的开销是 b,那么对于每个样本前向传播加上后向反馈的时间是 $(a+b)L$. 因为 a 和 b 在底层计算的时候可以当作常数,所以整体的时间复杂度是 $O(LN)$.

用卷积神经网络、对抗生成网络等模型来进行特征选择的优势非常明显,一方面网络的架构组织灵活,规模可以很庞大,而且 Caffe[70],Tensorflow[71] 这些深度学习的代码库也大大降低了神经网络进行机器学习、特征提取的使用难度;另一方面就在于特征提取和特征选择是自动的,在模型框架搭建好之后,训练过程中,模型能自动地挖掘特征,这实现了一定的智能. 但是神经网络用于特征选择也有很明显的缺陷,一方面特征是自动地被提取出来,可解释性和人类可读性比较差,模型的原理类似于黑盒,另一方面就是硬件资源的要求较高,算法难以部署在低配置硬件上. 神经网络除了可以用于文本分类、图像分类等监督学习任务上,也可以用于非监督学习或者半监督学习. 比如:去噪自编码器可以自动地提取出信号内在的特征,从而去除噪声[72],比如:用神经网络进行预训练的方法来实现半监督学习[73]. 除了对抗生成网络外,神经网络还有一个很重要的发展方向,就是迁移学习[74],即将一个领域学习到的网络结构和参数迁移到另一个领域,或者将大数据样本上学习到的特征表示迁移到小数据样本上. 因为神经网络的灵活性,使得用神经网络实现迁移学习变得比较容易.

5 利用主成分分析进行降维

主成分分析[75]的英文缩写为主成分分析(PCA, principal component analysis),是一种简单有效的降维方法. 一般的特征选择考虑的是自变量 X 与响应变量 Y 之间的联系,从而对 X 进行缩减,减小其维度,而 PCA 则是从 X 自身出发,来进行维度的缩减. 从这个角度上看,PCA 的过程中进行特征的提取和选择是无监督的,利用的是自变量本身的相关关系.

PCA 主要考虑分量中方差所包含的信息量,它认为方差越大,所包含的信息越多. 基于这个原则,对于一个矩阵提取主要成分的步骤包括5步:第1)步,计算出协方差矩阵 $\boldsymbol{A}$;第2)步,求出矩阵 $\boldsymbol{A}$ 的所有特征值和相应的单位正交特征向量;第3)步,根据方差越大包含信息越多的原则对特征值进行排序,选取那些值较大的特征;第4)步,计算主成分的载荷,即主成分与原变量之间的关联程度;第5)步,计算选取主成分之后的得分. 通过这5步,就可以得到合适的主成分,即所需要的特征. 在主成分分析的过程中,通常采用奇异值分解[76](SVD, singular-value decomposition)技术,因为奇异值分解是一种直观、计算简单、高效的求解矩阵的伪逆的技术. 从整体上看,矩阵的规模变小了,选出来的特征就是方差大的特征.

PCA 技术已经被广泛地应用于机器学习领域,比如 Sukthankar 等[77]开发了基于 PCA 技术的尺度不变特征转换(SIFT, scale-invariant feature transform)图像变换,Patil 等[78]利用层次 PCA 进行图像模糊处理, Berry 将 PCA 技术用于文本挖掘[79] PCA 的时间复杂度计算如下:记 N 为训练样本数量,F 为特征数量,那么 PCA 的协方差矩阵计算的开销为 $O(F^2N)$,特征值分解的计算开销为 $O(F^3)$,因此整体的时间复杂度为 $O(F^2N+F^3)$.

6 展望

随着多领域数据资源的整合,单一的特征选择方法往往不能满足全方位的数据挖掘需求. 因此,一种做法是:一种特征选择方法,或者一种机器学习流程被嵌入整个庞大的数据挖掘系统中,每个特征选择模块在特殊的数据处理 Pipeline 中发挥其特有的作用;另一种做法是:特征选择模型被做成可插拔的模块,随时可以被替换,从而满足新领域的数据需求. 这就要求每个模块要有足够

的鲁棒性和可移植性.

从特征选择算法本身的发展趋势来看,主要包括 3 个方面:1) 对抗神经网络以及其他深度学习框架将被广泛使用于各个特征选择的领域,带来新的机器学习应用;2) 迁移学习将大大缩短特征选择的周期和难度;3) 各种机器学习算法吸收各种特征选择算法的优点,将特征选择变得透明.

在实际的使用中,应该具体分析问题和数据规模,选取资源和效果在承受范围内的几种方案,通过模型训练,从中挑选出最合适的特征选择方案. 当现有的方案都不能满足需求时,需要参考已有方案思路,设计新的特征选择方案.

笔者分别从优点、缺点和时间复杂度 3 个方面,比较了五大类特征选择算法,如表 1 所示.

表 1　五类特征选择算法的比较

算法	优点	缺点	时间复杂度
相关性度量	计算简单	刻画的变量关系有限	$O(NF)$或$O(2^F N)$
Lasso 稀疏选择	可以将高维特征变得很稀释	当特征数量不是特别多的时候没有必要使用 Lasso	$O(NF^2)$
集成方法	面向模型最终的性能. 相对于相关性度量方法更加容易是全局最优. 模型训练过程中进行特征选择,不需要额外的代码,对于大规模数据很友好	实现较复杂	$O(MFN\log N)$
神经网络	自动提取和选择特征. 适用于监督学习以外的学习任务,也适合于迁移学习任务	特征的可解释性不强. 在训练上容易遇到问题. 原理上的数学证明的文献也很缺乏. 非常消耗硬件资源.	$O(LN)$
主成分分析	不需要考虑自变量和响应变量的关系. 提取出来的特征和自变量本身有关,适用于不同的学习任务	对于某些任务的效果,可能不如集成方法,或者神经网络等面向结果的方案. 当数据量变得特别大的时候,矩阵存储可能遇到问题.	$O(F^2N + F^3)$

7　结束语

在实际应用中,特征数量众多,可能存在不相关的特征,也可能存在相互依赖的关系,这在分类和预测时,容易导致训练模型所需的时间长、模型复杂、推广能力下降等. 而特征选择能剔除不相关和多余的特征,从而达到减少特征个数,提高模型精确度,减少运行时间的目的. 特征选择也叫属性选择. 特征选择方法有 3 个思路:1) Filter:对每一维特征“打分”,即给每一维特征赋予权重,然后依据权重排序;2) Wrapper:将特征选择看作一个搜索寻优问题,生成不同的组合,对组合进行评价;3) Embedded:在确定模型的过程中,挑选出那些对模型的训练有重要意义的属性.

笔者深入探讨了特征选择的五大类算法,分析了各种算法的解决思路、研究点、适用环境. 下一步将研究深度学习提取的特征,再结合机器学习的特征,进行分类研究;数据样本不足情况下的特征选择;迁移学习进行特征选择问题.

参考文献:

[1] Churchland P S, Sejnowski T J. The computational brain [M]. Cambridge: MIT Press, 2016: 1-120.

[2] Rizzo G, Troncy R, Hellmann S, et al. NERD meets NIF: lifting NLP extraction results to the linked data cloud[EB/OL]. Lyon: LDOW, 2012[2017-4-21]. http://www.eurecom.fr/fr/publication/3675/download/mm-publi-3675.pdf

[3] Koehn P, Hoang H, Birch A, et al. Moses: open source toolkit for statistical machine translation[C]//Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Prague: Association for Computational Linguistics, 2007: 177-180.

[4] Pastor G C, Mitkov R, Afzal N, et al. Translation universals: do they exist? a corpus-based NLP study of convergence and simplification[C]//8th AMTA Conference. Hawaii: Arts and Humanities Language and Linguistics, 2008: 75-81.

[5] Bernard D E. Multimodal natural language query system and architecture for processing voice and proximity-based queries: U.S. 376. 645[P]. 2008-05-20.

[6] Bernard D E. Multimodal natural language query system for processing and analyzing voice and proximity-based queries: U.S. 873. 654[P]. 2011-01-18.

[7] 王序文, 王小捷, 孙月萍. 双语主题跨语言伪相关反馈[J]. 北京邮电大学学报, 2013, 36(4): 81-84.

Wang Xuwen, Wang Xiaojie, Sun Yueping. Cross-lingual pseudo relevance feedback based on bilingual topics[J]. Journal of Beijing University of Posts and Telecommunications, 2013, 36(4): 81-84.

[8] Pedersen T, Pakhomov S V S, Patwardhan S, et al. Measures of semantic similarity and relatedness in the biomedical domain[J]. Journal of Biomedical Informatics, 2007, 40(3): 288-299.

[9] 赵学军, 李育珍, 雷书彧. 基于遗传算法优化的稀疏表示图像融合算法[J]. 北京邮电大学学报, 2016, 39(2): 73-76, 87.
Zhao Xuejun, Li Yuzhen, Lei Shuyu. An image fusion method with sparse representation based on genetic algorithm optimization[J]. Journal of Beijing University of Posts and Telecommunications, 2016, 39(2): 73-76, 87.

[10] Sonka M, Hlavac V, Boyle R. Image processing, analysis, and machine vision[M]. Boston: Cengage Learning, 2014: 312-389.

[11] Akata Z, Perronnin F, Harchaoui Z, et al. Good practice in large-scale learning for image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 507-520.

[12] 李亚, 王广润, 王青. 基于深度卷积神经网络的跨年龄人脸识别[J]. 北京邮电大学学报, 2017, 40(1): 84-88, 110.
Li Ya, Wang Guangrun, Wang Qing. A deep joint learning approach for age invariant face verification[J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40(1): 84-88, 110.

[13] Wang S J, Chen H L, Yan W J, et al. Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine [J]. Neural Processing Letters, 2014, 39(1): 25-43.

[14] Sridhar S, Oulasvirta A, Theobalt C. Interactive markerless articulated hand motion tracking using RGB and depth data[C]// Proceedings of the IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 2456-2463.

[15] Behoora I, Tucker C S. Machine learning classification of design team members' body language patterns for real time emotional state detection[J]. Design Studies, 2015 (39): 100-127.

[16] Chen X, Zheng Z, Yu Q, et al. Web service recommendation via exploiting location and QoS information [J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(7): 1913-1924.

[17] Skowron P, Faliszewski P, Lang J. Finding a collective set of items: from proportional multirepresentation to group recommendation [J]. Artificial Intelligence, 2016 (241): 191-216.

[18] Broder A, Fontoura M, Josifovski V, et al. A semantic approach to contextual advertising[C]// Proceedings of the 30th annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Amsterdam: ACM, 2007: 559-566.

[19] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques [C]// Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2002: 79-86.

[20] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: a deep learning approach[C]// Proceedings of the 28th International Conference on Machine Learning (ICML-11). Boston: IMLS, 2011: 513-520.

[21] Daum F, Huang J. Curse of dimensionality and particle filters[C]// Aerospace Conference, 2003. Proceedings. 2003 IEEE. Montana: IEEE, 2003: 1979-1993.

[22] Hawkins D M. The problem of overfitting[J]. Journal of Chemical Information and Computer Sciences, 2004, 44 (1): 1-12.

[23] Domingos P. A few useful things to know about machine learning[J]. Communications of the ACM, 2012, 55 (10): 78-87.

[24] 周志华. 机器学习 [M]. 北京:清华大学出版社, 2016: 247-266.

[25] Xing E P, Jordan M I, Karp R M. Feature selection for high-dimensional genomic microarray data[C]// ICML. Williamstown: ICML, 2001: 601-608.

[26] Singh S R, Murthy H A, Gonsalves T A. Feature selection for text classification based on gini coefficient of inequality[J]. FSDM, 2010(10): 76-85.

[27] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization[C]// ICML. Nashville: ICML, 1997: 412-420.

[28] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (8): 1226-1238.

[29] Witten I H, Frank E, Hall M A, et al. Data mining: practical machine learning tools and techniques [M].

Burlington：Morgan Kaufmann，2016.

[30] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological)，1996，58(1)：267-288.

[31] Hoerl A E，Kennard R W. Ridge regression：biased estimation for nonorthogonal problems [J]. Technometrics，1970，12(1)：55-67.

[32] Murphy K P. Machine learning：a probabilistic perspective [M]. Cambridge：MIT Press，2012.

[33] Zou H，Hastie T. Regularization and variable selection via the elastic net [J]. Journal of the Royal Statistical Society：Series B (Statistical Methodology)，2005，67(2)：301-320.

[34] Zhou Y，Jin R，Hoi S. Exclusive lasso for multi-task feature selection[C]// International Conference on Artificial Intelligence and Statistics. Sardinia，Italy：[s. n.]，2010：988-995.

[35] Efron B，Hastie T，Johnstone I，et al. Least angle regression[J]. The Annals of Statistics，2004，32(2)：407-499.

[36] Sun Y，Chen Y，Wang X，et al. Deep learning face representation by joint identification-verification[C]// Advances in Neural Information Processing Systems. Montreal：NIPS，2014：1988-1996.

[37] Breiman L. Bagging predictors[J]. Machine Learning，1996，24(2)：123-140.

[38] Breiman L. Random forests [J]. Machine Learning，2001，45(1)：5-32.

[39] Rätsch G，Onoda T，Müller K R. Soft margins for adaBoost[J]. Machine Learning，2001，42(3)：287-320.

[40] De'Ath G. Boosted trees for ecological modeling and prediction[J]. Ecology，2007，88(1)：243-251.

[41] Chen T，Guestrin C. Xgboost：a scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco：ACM，2016：785-794.

[42] Burrows W R，Benjamin M，Beauchamp S，et al. CART decision-tree statistical analysis and prediction of summer season maximum surface ozone for the Vancouver，Montreal，and Atlantic regions of Canada [J]. Journal of Applied Meteorology，1995，34(8)：1848-1862.

[43] Menze B H，Kelm B M，Masuch R，et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data[J]. BMC Bioinformatics，2009，10(1)：213.

[44] Díaz-Uriarte R，De Andres S A. Gene selection and classification of microarray data using random forest[J]. BMC Bioinformatics，2006，7(1)：3.

[45] Pal M. Random forest classifier for remote sensing classification[J]. International Journal of Remote Sensing，2005，26(1)：217-222.

[46] Chen X W，Liu M. Prediction of protein-protein interactions using random decision forest framework[J]. Bioinformatics，2005，21(24)：4394-4400.

[47] Liu G，Nguyen T T，Zhao G，et al. Repeat buyer prediction for e-commerce[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco：ACM，2016：155-164.

[48] Volkovs M. Two-stage approach to item recommendation from user sessions[C]// Proceedings of the 2015 International ACM Recommender Systems Challenge. Vienna：ACM，2015：3.

[49] Hinton G E，Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science，2006，313(5786)：504-507.

[50] LeCun Y，Bengio Y，Hinton G. Deep learning[J]. Nature，2015，521(7553)：436-444.

[51] Deng L，Yu D. Deep learning：methods and applications[J]. Foundations and Trends® in Signal Processing，2014，7(3-4)：197-387.

[52] Krizhevsky A，Sutskever I，Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. Stateline：2012：1097-1105.

[53] Kim Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha：Association for Computational Linguistics，2014：1746-1751.

[54] Ji S，Xu W，Yang M，et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2013，35(1)：221-231.

[55] Graves A，Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks，2005，18(5)：602-610.

[56] Sukhbaatar S，Weston J，Fergus R. End-to-end memory networks[C]// Advances in Neural Information Processing Systems. Montreal：IEEE，2015：2440-2448.

[57] Wen T H，Gasic M，Mrksic N，et al. Semantically con-

ditioned lstm-based natural language generation for spoken dialogue systems[EB/OL]. Issa card City: arXiv Preprint arXiv, 2015[2017-3-12]. https://arxiv.org/abs/1503.00075.

[58] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: [s.n.], 2016: 770-778.

[59] Hinton G E. To recognize shapes, first learn to generate images[J]. Progress in Brain Research, 2007(165): 535-547.

[60] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv Preprint arXiv:1511.06434, 2015.

[61] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [J]. arXiv Preprint arXiv:1511.06434, 2015.

[62] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[63] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]// Advances in Neural Information Processing Systems. Stateline: [s.n.], 2012: 1097-1105.

[64] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1-9.

[65] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2016: 770-778.

[66] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. Montreal: [s.n.], 2014: 2672-2680.

[67] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[EB/OL]. Issa card City: arXiv Preprint arXiv, 2016 [2017-2-3]. https://arxiv.org/abs/1609.04802, 2016.

[68] Reed S, Akata Z, Yan X, et al. Generative adversarial text to image synthesis[C]// Proceedings of The 33rd International Conference on Machine Learning. New York: [s.n.], 2016: 3.

[69] Li J, Monroe W, Shi T, et al. Adversarial learning for neural dialogue generation[EB/OL]. Issa card City: arXiv Preprint arXiv,2017[2017-5-15]. https://arxiv.org/abs/1701.06547.

[70] Jia Y, Shelhamer E, Donahue J, et al. Caffe: convolutional architecture for fast feature embedding[C]// Proceedings of the 22nd ACM International Conference on Multimedia. Orlando: ACM, 2014: 675-678.

[71] Abadi M, Agarwal A, Barham P, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems[EB/OL]. Issa card City: arXiv Preprint arXiv, 2016 [2017-5-10]. https://arxiv.org/abs/1603.04467.

[72] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the 25th International Conference on Machine Learning. Helsinki : ACM, 2008: 1096-1103.

[73] Erhan D, Bengio Y, Courville A, et al. Why does unsupervised pre-training help deep learning? [J]. Journal of Machine Learning Research, 2010, 11(2): 625-660.

[74] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[75] Dunteman G H. Principal components analysis[M]. New York: Sage Publications, 1989: 31-70.

[76] Klema V, Laub A. The singular value decomposition: its computation and some applications[J]. IEEE Transactions on Automatic Control, 1980, 25(2): 164-176.

[77] Ke Y, Sukthankar R. PCA-SIFT: a more distinctive representation for local image descriptors[C]// Computer Vision and Pattern Recognition, 2004, CVPR 2004 [C]// Proceedings of the 2004 IEEE Computer Society Conference on. Washington: IEEE, 2004: 506-513.

[78] Patil U, Mudengudi U. Image fusion using hierarchical PCA[C]// Image Information Processing (ICIIP), 2011 International Conference on. Shimla: IEEE, 2011: 1-6.

[79] Berry M W. Survey of text mining[J]. Computing Reviews, 2004, 45(9): 548.