文章编号:1007－6735(2020)03－0247－06 DOI: 10.13255/j.cnki.jusst.20190318001

云环境中大数据挖掘的有效花费研究

朱小栋， 徐 怡， 魏紫钰
（上海理工大学 管理学院，上海 200093）

摘要： 为平衡云计算资源的租用量与云环境中数据挖掘的计算结果准确率，得到最优的性价比，以监督式学习的卷积神经网络（CNN）为例，探究了 CNN 迭代次数与准确率的演化规律。选择经典图像数据集 MNIST 和文本数据集 IMDB 作为代表展开实验，发现在不同类型的数据集中，当 CNN 迭代接近最优解时，每提高很小的准确率，耗费的机时陡增，称之为长尾现象。验证在真实云环境中，当大数据挖掘的长尾现象发生且满足企业准确率需求的情况下，选择提前结束取代最高精度时结束，均可以节省大量云资源成本。研究结果对于合理运用云计算资源，降低云服务租用成本，具有实用价值与现实意义。

关键词： 云计算资源；有效花费；卷积神经网络；长尾现象
中图分类号： TP 39 **文献标志码：** A

Effective cost of big data mining in cloud environment

ZHU Xiaodong, XU Yi, WEI Ziyu
(*Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: In order to balance the renting quantity of cloud computing resources and the accuracy of data mining in cloud, the optimum cost performance ratio is obtained. Taking the convolution neural network (CNN) as an example, the evolution patterns of the number of iterations and accuracy of CNN was explored. A lot of experiments were performed upon the image dataset MNIST and the text dataset IMDB. The results show that in different types of data sets, the machine time consumed increases sharply with a small increase in accuracy when the optimal solution is approached, which is called the long tail phenomena. Correspondingly, in the real cloud environment, when the long tail phenomenon of big data mining occurs and the accuracy is satisfied, terminating the performance of CNN in cloud in advance rather than at the convergence time can save a lot of cloud resource costs. The results have practical value and practical significance for the rational use of cloud computing resources and the reduction of cloud rental cost.

Keywords: *cloud computing resources*; *effective cost*; *convolutional neural network*; *long tail phenomenon*

收稿日期: 2019−03−18
基金项目: 国家自然科学基金资助项目（71771152）；上海市人民政府发展研究中心“基于互联网+的上海创新发展”研究基地决策咨询研究项目（2019-YJ-L04-A）；2020 上海市教委高校智库内涵建设项目
第一作者: 朱小栋（1981−），男，副教授．研究方向：大数据管理、数据挖掘、云计算、信息安全．E-mail: zhuxd@usst.edu.cn

随着互联网与通信技术的迅猛发展，我们已经处于数据爆炸的时代[1]。例如，脸谱网每月约有60亿张新照片，YouTube每分钟约400 h视频被上传。2010年时中国网页规模就已达到600亿个，年增长率达到78.6%[2]。这种爆炸性增长的数据推动了大范围的数据挖掘，如商业、政府、医疗保健等。所以，相应地，大数据的有效分析和挖掘具有不可估量的经济价值。

大数据时代来临的同时，大多数数据挖掘算法在计算复杂度上呈指数级增长，数据挖掘过程要花费数小时甚至数天才能完成。因此，大数据挖掘常常需要大量的计算资源与计算空间。然而，许多企业，尤其是中小企业和组织无法负担大型数据挖掘[3]。就像加州大学伯克利分校可靠自适应分布式系统实验室的Armbrust[4]提到的那样："大规模并行任务处理能够获得与传统计算累加相似的效率，使用一千台服务器运算一个小时的成本，与一台服务器计算一千个小时不相上下。这种资源的弹性是史无前例的——它意味着用户不必为扩展花费过多的成本。"所以，云计算资源的租用成为有效的途径[5]，特别对于中小企业而言，节省了大量的开支与精力。

但是，云计算资源也在随着大数据的兴起而愈加昂贵。如果对云计算资源进行计算，不适当地管理与购买计算资源，那么利用云计算资源(即计算成本)的成本会高得出奇[6]。DrawerKVM公司在2018年对欧洲进行调查，过去3年英国企业在云计算上的平均支出为360 000欧元，而数据中心的平均支出为343 000欧元，未来3年，英国的云计算支出预计将增长37%。Claranet总经理罗伯特说，英国企业对云的需求很大[7]。不止英国，全世界的企业甚至个人在云计算中的花费都愈加增长。以Amazon EC2为例，其中型(Medium)虚拟机(3.75 GB内存，2 ECU计算单元，410 GB存储，0.16美元/h)的配置是小型(Small)虚拟机(1.7 GB内存，1 ECU计算单元，160 GB存储，0.08美元/h)的两倍，其价格也是小型虚拟机的两倍。运行100个M4超大亚马逊EC2虚拟机(VM)的情况下，费用为每天583美元。所以，云计算资源成为热流的同时，如何节省租金更加高效地使用云资源成为又一难题[8]。

很多企业在使用云资源运算时，有时并不需要100%的准确率。例如一些电商企业，会对大量的用户与产品进行数据挖掘，以此得知哪些产品更受哪些用户喜欢，哪些产品可以进行相应促销等等。在这个过程中，适当幅度的误差是可以接受的，营销人员可以根据大致的画像就可以作出相应的决断，而并不需要100%的精确信息。另外，事实上，在一些数据挖掘实例中，并不会得到完全精准的预测结果，例如天气的实时预报和零售客户的细分等等。

本文的研究目的是如何能以最低的花费获得足够的云计算资源。本文以当前热门的数据挖掘算法CNN为例进行研究，发现CNN在迭代过程中的长尾现象，即在迭代过程中，当CNN的计算准确率为0.90左右时，有一个突进现象的存在。相应地，当长尾现象发生时，满足企业需求的准确率可能只需要CNN运算能够达到的最大准确率成本的3.34%。即假如CNN能够达到0.993 7的准确率，但0.885 7的准确率就能满足企业需求，并且0.885 7的准确率所需的成本可能只是0.993 7准确率所需成本的3.34%。

1 云计算的消费模型

1.1 云计算

关于云计算的定义，截至目前为止仍然没有一个统一的结果[9]，这些定义既有从技术角度进行的描述，也有从商业角度进行的总结定义，但是总体来说，云计算可从"服务"和"平台"两个角度去考虑，即云计算包含云计算平台和云计算服务这两个概念。云计算服务指的是一种新型的商业模式，旨在给用户提供可靠的在线服务。而云计算平台是伴随云计算服务应运而生的，更像是一种操作系统，通过一些技术手段将分布在各地的计算机以网络的方式进行连接，并在逻辑上以整体的方式呈现。云计算平台和云计算服务的关系就如同底层建筑和上层建筑，但是这两者之间也没有必然的对应关系[10]。

就本文的研究问题而言，本文重点在于云计算服务方面，但是云计算服务也是要以云计算平台为支撑，只有二者完美结合，才能实现为大规模的用户提供稳定、可靠、低成本的服务。但是，即使出现云计算服务这么优势明显的商业模式，仍然有一些企业，尤其是中小企业负担不起云计算的服务费用，这成为笔者一直重点关注的问题。

1.2　消费模型

云计算平台已经层出不穷，国外以亚马逊为例，国内以阿里云和华为云为例介绍云计算资源的计费模式，详见表 1。

表 1　三大厂商计费模式

Tab.1　Three major vendors' billing model

计费模式	厂商		
	Amazon EC2	阿里云	华为云
按需定价	√	√	√
预留实例	√		
Spot 实例	√		
包年包月		√	√
竞价型计费		√	
阶梯计价		√	

从以上三大厂商的计费模式可以看出，三大厂商都有按需计费这一消费模式。按需计费指的是用户可以根据运行的实例以按小时或按分钟甚至可以支持按秒的方式为计算容量付费。而无需签订长期合同或支付预付款。这种方式比较灵活，可灵活控制成本。

在本文研究中，为了提升泛化能力，更集中突出所研究的问题，以按需计费方式和 CNN 为例进行研究，见式(1)。

$$C = PT \tag{1}$$

式中：C表示所需要的所有花费；P表示单价，即每秒所需花费；T为 CNN 计算时所需时间，s。

事实上，为了更加集中于所研究的问题，只计算在运算过程中的花费，而忽略了存储、迁移等方面所需的云计算资源的花费。

2　卷积神经网络相关概述

2.1　卷积神经网络

卷积神经网络(CNN)是神经网络的一种，是一种学习效率很高的深度学习模型，对于很多模式识别领域尤其是图像识别方面都取得了良好的识别效果[11]，LeCun 曾经提出的对于手写数字识别的 CNN 模型 LeNet-5[12]结构，就具有极高的准确率。

CNN 的基本结构由输入层、卷积层、池化层(也称下采样层)、全连接层和输出层组成。卷积层和池化层一般根据所需情况取若干个，交替设置。卷积神经网络含有最突出的 3 个特点，即局部连接、权值共享和池化操作，有效地降低了网络的复杂度，减少了训练参数的数量，降低特征维度并且改善结果。

2.2　准确度估计

卷积神经网络是一种监督式学习的神经网络[13]，由于知道原先的分类，故可以准确计算出利用卷积神经网络进行分类后的精确性。

对于经典的二分类问题，真正例(true positives，TP)是指实际上是正例的标记为正例；假正例(false positives，FP)是指实际上是反例的数据被标记为正例；真反例(true negatives，TN)是指实际上是反例的数据被标记为反例；假反例(false negatives，FN)是指实际上是正例的数据被标记为反例[14-15]。

准确率反映了分类模型对整个样本的判定能力，定义见式(2)。

$$A = \frac{n_{\mathrm{TP}} + n_{\mathrm{TN}}}{n_{\mathrm{TP}} + n_{\mathrm{FN}} + n_{\mathrm{FP}} + n_{\mathrm{TN}}} \tag{2}$$

式中：A 为准确率；n_{TP}，n_{FP}，n_{TN}，n_{FN} 分别表示算法在数据集上的各种测试结果。

分类问题中更关注的是准确率，这是一个比均方损失或者交叉熵损失更重要的量度。在这里，主要利用准确率评估分类的准确性。

3　实　验

3.1　实验设置

实验在 Intel(R) Core(TM) i5-4210U CPU @ 1.70 GHz 2.40 GHz 下,利用 Python3.5.2 Keras 完成，数据集利用 CNN 常用的手写字体数据集 MNIST 和电影评论情感分类数据集 IMDB。

案例研究程序包含以下几个步骤：

a. 数据集准备。准备要实验的数据集。

b. 数据分类。利用 CNN 对数据集进行实验分类，并确定准确率。

c. 准确率–时间比较。对于每一组实验，通过算法在每次迭代中获得的准确率，与按需模型的每次迭代结束时算法所花费的计算时间一起示出。

d. 分析和讨论。比较结果进行分析和讨论。

3.2　数据集准备

MNIST 手写数据集是深度学习最常用的数据

集之一，是美国国家标准与技术研究所（National Institute of Standards and Technology，NIST）所提出的。训练集由来自 250 个不同人手写的数字构成，其中 50% 是高中学生，50% 来自人口普查局的工作人员。测试集也是同样比例的手写数字数据。图 1 为 MNIST 数据集的可视化样例。

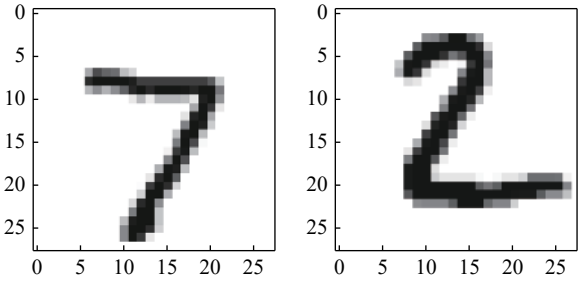


图 1　MNIST 数据集可视化样例

Fig.1　MNIST dataset visualization example

IMDB 影评情感数据也是作文本情感分类常用的数据集之一，是斯担福大学人工智能实验室整理的一套 IMDB 影评的情感数据[16]。

3.3　实验结果

实验分别采用 5 层、7 层和 9 层的 CNN 对 MNIST 和 IMDB 数据集进行分类，相应的数据集分布见表 2。

表 2　数据集分布表

Tab.2　Dataset distribution table

数据集	训练样本	测试样本
MNIST	60 000	10 000
IMDB	25 000	25 000

利用 MNIST 进行实验时，采用二维卷积层，卷积核的数目为 32，卷积核的大小为 3 × 3，损失函数采用交叉熵，优化器采用 Adadelta。从图 2 中可以看出，深度 K=5，7，9 时，训练集的准确率总有一个突进，然后再缓缓趋向平稳，这称之为 CNN 迭代中的长尾现象。其实不止 CNN，只要带有迭代性质的算法基本上都会有长尾现象的出现，例如 K-means[17]，而这些算法在实际运用中都普遍使用，因此本研究也具有非常重要的普适意义。

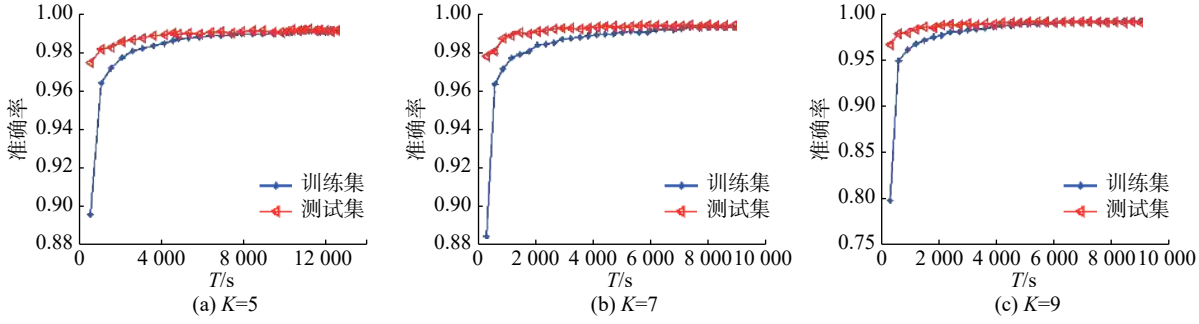


图 2　MNIST 训练集和测试集在准确率和时间之间的相关性

Fig.2　Relationship between the time and accuracy for the training and test data of MNIST

以华为云为例，按需计费开通弹性云服务器（elastic cloud server）实例，购买规格为 h1.2xlarge.4 |8 核| 32 G，系统盘为 40 G 时，都有统一的计价标准，即只要选定固定的规格，就会有固定的价格。所以根据之前确定花费的模型可知，决定训练成本的只是卷积神经网络完成迭代需要花费的时间。因此，成本花费表也就是计算时间表，具体数据见表 3。

由表 3 可以看出，以 K=5 为例，当准确率为 0.896 7 时，计算时间为 549 s；当准确率为 0.990 1 时，计算时间为 788 7 s；当准确率为 0.992 1 时，计算时间已经达到 12 222 s，是准确率 0.896 7 所需计算时间的 22.26 倍，是准确率 0.990 1 所需计算时间的 1.55 倍。也就是说，如果要求准确率接近为 0.9 时，只需要准确率接近为 0.99 的 6.96% 花费，只需要准确率为 0.992 1 的 4.49% 花费即可。同样，K=7 时，如果要求准确率为 0.885 7，只需要准确率为 0.99 的 6.70% 花费，只需要准确率为 0.993 7 的 3.34% 花费即可。K=9 时，如果要求准确率为 0.799 4，只需要准确率接近为 0.99 的 5.41% 花费，只需要准确率为 0.993 3 的 3.43% 花费即可。

表 3　MNIST 计算时间表

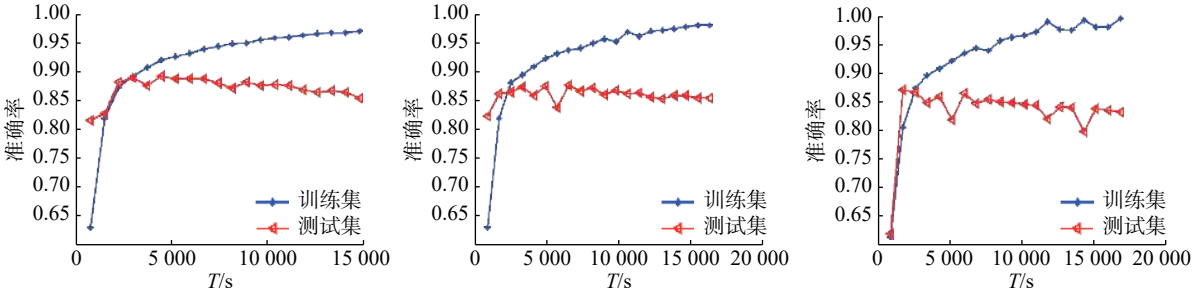
Tab.3　MNIST calculation time

K	时间/s	比例 1/%	比例 2/%	准确率
	549	4.49	6.96	0.896 7
5	7 887	64.53	100.00	0.990 1
	12 222	100.00	–	0.992 1
	295	3.34	6.70	0.885 7
7	4 402	49.91	100.00	0.990 0
	8 820	100.00	–	0.993 7
	307	3.43	5.41	0.799 4
9	5 670	63.29	100.00	0.990 2
	8 959	100.00	–	0.993 3

由实验可以看出，云计算的花费随着准确率增加会有一个爆发式的增长，接下来的增长就会特别缓慢。企业或个人在某些时候并不需要非常精准的计算效果。在此实验下，满足企业需求的准确率所需要的成本，可能只需要 CNN 运算能够达到的最大准确率的成本的 3.34% ～4.49%。

为了避免偶然性，故又采取了另一个文本数据集 IMDB，同样还是利用 CNN 进行实验。

利用 IMDB 数据集进行实验时，采用一维卷积层，卷积核数目为 128，卷积核大小为 3，损失函数采用对数函数，优化器采用 rmsprop。同样地，从图 3 可以看出，即使采用 CNN 所不擅长训练的文本数据也有长尾现象的存在。由表 4 可以看出：当 K=5 时，如果要求准确率为 0.892 7，只需要准确率为 0.950 1 的 33.43% 花费，或准确率为 0.969 7 的 20.26% 花费即可；当 K=7 时，如果要求准确率为 0.895 9，只需要准确率为 0.950 9 的 39.91% 花费，或准确率为 0.982 1 的 21.00% 花费即可；当 K=9 时，如果要求准确率为 0.896 7，只需要准确率为 0.957 2 的 40.24% 花费，或准确率为 0.996 6 的 20.25% 花费即可。所以在此实验下，满足企业准确率要求所需要的成本，可能只需要 CNN 运算能够达到的最大准确率成本的 20.25%～21.00%。



(a) K=5　(b) K=7　(c) K=9

图 3　IMDB 训练集和测试集在准确率和时间之间的相关性

Fig.3　Relationship between the time and accuracy for the training and test data of IMDB

表 4　IMDB 计算时间表

Tab.4　IMDB calculation time

K	时间/s	比例 1/%	比例 2/%	准确率
	2 949	20.20	33.43	0.892 7
5	8 822	60.01	100.00	0.950 1
	14 701	100.00	–	0.969 7
	3 244	21.00	39.91	0.895 9
7	8 129	52.61	100.00	0.950 9
	15 450	100.00	–	0.982 1
	3 397	20.25	40.24	0.896 7
9	8 441	50.31	100.00	0.957 2
	16 777	100.00	–	0.996 6

另外，从图 3 也可以看出，测试集出现过明显的波动，这是因为 CNN 在文本处理方面效果并不稳定。但从测试集的准确率曲线可以得出，并不是训练准确率越高，测试集的准确率也越高，所以最高的训练准确率有时并不是最优的选择。总的来说，企业只需要选择自己所需要的合适准确率，并不用一味追求最高的准确率，这样才能达到经济效益的最大化。

4　结　论

以卷积神经网络的长尾现象为切入点，揭示了云环境中有效花费的问题，为企业如何合理化

运用云资源作出参考。以 CNN 为例，探索了收敛特征的数据挖掘算法在迭代过程中存在长尾现象，进而推广到大数据的数据挖掘上。当长尾现象发生且满足企业准确率要求的情况下，提前结束数据挖掘算法的运行可以为企业节省大量云计算成本。因此，企业可以小得多的成本来得到其所需要的准确率，而不需要耗费不必要的代价来租用过量的云资源。在大数据挖掘的背景下，企业租用云资源成本会急剧增加，不必要的花费会给企业带来沉重的负担，本研究为企业降低云资源租用成本提供了思路。在实际情况中，企业需要考虑的地方还包括存储、迁移时需要的云资源，这将是本研究下一步的工作。

参考文献：

[1] 王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望[J]. 计算机学报, 2013, 36(6): 1125–1138.

[2] 于戈, 谷峪, 鲍玉斌, 等. 云计算环境下的大规模图数据处理技术[J]. 计算机学报, 2011, 34(10): 1753–1767.

[3] 江涛. 当管理遭遇“云”——云计算: 不只是节省成本[J]. 管理学家: 实践版, 2012(5): 30–38.

[4] ARMBRUST M. Above the clouds: a berkeley view of cloud computing[J]. Science, 2009, 53(4): 50–58.

[5] 冯登国, 张敏, 张妍, 等. 云计算安全研究[J]. 软件学报, 2011, 22(1): 71–83.

[6] 陈秀惠. 如何避免云计算的成本超支[J]. 计算机与网络, 2018, 44(18): 38–40.

[7] ABADI M. TensorFlow: learning functions at scale[J]. ACM SIGPLAN Notices, 2016, 51(9): 1.

[8] 何雪锋, 陈静利, 张鑫. 基于人工智能、大数据和云计算的作业成本法探究——以我国烟草工业企业为例[J]. 财会月刊, 2018(17): 69–72.

[9] 张建勋, 古志民, 郑超. 云计算研究进展综述[J]. 计算机应用研究, 2010, 27(2): 429–433.

[10] 姚宏宇, 田溯宁. 云计算: Cloud computing: 大数据时代的系统工程[M]. 北京: 电子工业出版社, 2013.

[11] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229–1251.

[12] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[13] HAYKIN S. 神经网络原理[M]. 2 版. 北京: 机械工业出版社, 2006.

[14] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[15] 周志华, 王珏. 机器学习及其应用[M]. 北京: 清华大学出版社, 2007.

[16] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACM, 2011: 142–150.

[17] HE Q, ZHU X D, LI D W, et al. Cost-effective big data mining in the cloud: a case study with K-means[C]// Proceedings of the 2017 IEEE 10th International Conference on Cloud Computing. Honolulu: IEEE, 2017: 74–81.

（编辑：丁红艺）