**内蒙古财经大学职业学院高职毕业论文（毕业设计）**

**现代信息技术在会计中的应用构想**

内容提要

随着云计算，大数据，机器学习，各种硬件设施的发展完善，机器学习算法应用于各个行业的时机也逐渐趋于成熟。2016年3月15日，谷歌公司开发的人工智能机器AlphaGo以总比分4:1战胜围棋世界冠军李世石，轰动世界的“人机大战”落下帷幕。2017年1月4日，升级为Master的AlphaGo，经过7天的“踢馆”大战，力克国际60位顶级围棋高手（其中包括16位世界冠军），再次掀起“腥风血雨”。事实上，早在20世纪50年代，人工智能便开始向人类发起挑战。当时来自IBM工程研究组的萨缪尔（Samuel）开发出一款跳棋程序，该程序能够在与人对弈的过程中，不断累积经验提升棋艺，并于1959年战胜了萨缪尔本人。应该说，无论是半个多世纪前的“跳棋对决”，还是当前的“人机大战”，推动人工智能发展的核心动力未曾改变，即计算科学的璀璨明珠——机器学习。

关键词：机器学习算法 机器智能 大数据

目 录

[一、现代信息技术发展简介 2](#_Toc23466)

[（一）云计算 2](#_Toc14151)

[（二）大数据 2](#_Toc24175)

[（三）机器学习 2](#_Toc21742)

[（四）5G 2](#_Toc8126)

[（五）硬件设备发展 3](#_Toc7643)

[二、数据处理框架 4](#_Toc2956)

[（一）数据获取 4](#_Toc16207)

[（二）探索性数据分析与可视化 4](#_Toc20147)

[（三）特征预处理 4](#_Toc31985)

[（四）挖掘建模 4](#_Toc1959)

[三、现代信息技术在会计中的应用构想 5](#_Toc18928)

**现代信息技术在会计中的应用构想**

一、现代信息技术发展简介

**（一）云计算**

云计算（cloud computing）是分布式计算的一种，指的是通过网络“云”将巨大的数据计算处理程序分解成无数个小程序，然后，通过多部服务器组成的系统进行处理和分析这些小程序得到结果并返回给用户。云计算早期，简单地说，就是简单的分布式计算，解决任务分发，并进行计算结果的合并。因而，云计算又称为网格计算。通过这项技术，可以在很短的时间内（几秒种）完成对数以万计的数据的处理，从而达到强大的网络服务。

现阶段所说的云服务已经不单单是一种分布式计算，而是分布式计算、效用计算、负载均衡、并行计算、网络存储、热备份冗杂和虚拟化等计算机技术混合演进并跃升的结果。云计算的出现极大的促进了中小企业信息化进程，以前企业想要实现信息化，个人想要创业先得购买昂贵的服务器，现在只需要租用一台服务器便可以，并且省下了服务器维护费用，使用灵活，随你的使用情况灵活付费。

### （二）大数据

大数据（big data），IT行业术语，是指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。 

在[维克托·迈尔-舍恩伯格](https://baike.baidu.com/item/%E7%BB%B4%E5%85%8B%E6%89%98%C2%B7%E8%BF%88%E5%B0%94-%E8%88%8D%E6%81%A9%E4%BC%AF%E6%A0%BC" \t "/home/niangu/文档\\x/_blank)及肯尼斯·库克耶编写的《[大数据时代](https://baike.baidu.com/item/%E5%A4%A7%E6%95%B0%E6%8D%AE%E6%97%B6%E4%BB%A3/15434499" \t "/home/niangu/文档\\x/_blank)》   中大数据指不用随机分析法（[抽样调查](https://baike.baidu.com/item/%E6%8A%BD%E6%A0%B7%E8%B0%83%E6%9F%A5" \t "/home/niangu/文档\\x/_blank)）这样捷径，而采用所有数据进行分析处理。大数据的5V特点（IBM提出）：[Volume](https://baike.baidu.com/item/Volume/17610592" \t "/home/niangu/文档\\x/_blank)（大量）、[Velocity](https://baike.baidu.com/item/Velocity/1398152" \t "/home/niangu/文档\\x/_blank)（高速）、[Variety](https://baike.baidu.com/item/Variety/191328" \t "/home/niangu/文档\\x/_blank)（多样）、[Value](https://baike.baidu.com/item/Value/2285610" \t "/home/niangu/文档\\x/_blank)（低价值密度）、[Veracity](https://baike.baidu.com/item/Veracity/19362178" \t "/home/niangu/文档\\x/_blank)（真实性）。大数据是机器学习算法的基础，没有数据一切算法都是空谈。

**（三）机器学习**

机器学习是一门多领域交叉学科，涉及概率论、统计学、[逼近论](https://baike.baidu.com/item/%E9%80%BC%E8%BF%91%E8%AE%BA/967006" \t "/home/niangu/文档\\x/_blank)、[凸分析](https://baike.baidu.com/item/%E5%87%B8%E5%88%86%E6%9E%90" \t "/home/niangu/文档\\x/_blank)、[算法复杂度](https://baike.baidu.com/item/%E7%AE%97%E6%B3%95%E5%A4%8D%E6%9D%82%E5%BA%A6" \t "/home/niangu/文档\\x/_blank)理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。

它是[人工智能](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/9180" \t "/home/niangu/文档\\x/_blank)的核心，是使计算机具有智能的根本途径。

**（四）5G**

**第五代移动通信技术**（英语：5th generation mobile networks或5th generation wireless systems、5th-Generation，简称**5G**或**5G技术**）是最新一代[蜂窝移动通信](https://baike.baidu.com/item/%E8%9C%82%E7%AA%9D%E7%A7%BB%E5%8A%A8%E9%80%9A%E4%BF%A1/106306" \t "/home/niangu/文档\\x/_blank)技术，也是即[4G](https://baike.baidu.com/item/4G/523884" \t "/home/niangu/文档\\x/_blank)（[LTE-A](https://baike.baidu.com/item/LTE-A/8136979" \t "/home/niangu/文档\\x/_blank)、[WiMax](https://baike.baidu.com/item/WiMax/506486" \t "/home/niangu/文档\\x/_blank)）、[3G](https://baike.baidu.com/item/3G/99011" \t "/home/niangu/文档\\x/_blank)（[UMTS](https://baike.baidu.com/item/UMTS/1378941" \t "/home/niangu/文档\\x/_blank)、[LTE](https://baike.baidu.com/item/LTE/6439078" \t "/home/niangu/文档\\x/_blank)）和[2G](https://baike.baidu.com/item/2G/3110701" \t "/home/niangu/文档\\x/_blank)（[GSM](https://baike.baidu.com/item/GSM/210860" \t "/home/niangu/文档\\x/_blank)）系统之后的延伸。5G的性能目标是高数据速率、减少延迟、节省能源、降低成本、提高系统容量和大规模设备连接。Release-15中的5G规范的第一阶段是为了适应早期的商业部署。Release-16的第二阶段将于2020年4月完成，作为IMT-2020技术的候选提交给[国际电信联盟](https://baike.baidu.com/item/%E5%9B%BD%E9%99%85%E7%94%B5%E4%BF%A1%E8%81%94%E7%9B%9F/502493" \t "/home/niangu/文档\\x/_blank)（ITU） [1]  。ITU IMT-2020规范要求速度高达20 Gbit/s，可以实现宽信道带宽和大容量[MIMO](https://baike.baidu.com/item/MIMO/1881" \t "/home/niangu/文档\\x/_blank)。

2019年10月31日，三大运营商公布[5G商用](https://baike.baidu.com/item/5G%E5%95%86%E7%94%A8/24112622" \t "/home/niangu/文档\\x/_blank)套餐，并于11月1日正式上线5G商用套餐。5G技术的应用将会极大的促进信息的流通，硬件之间的远程巨量信息沟通的障碍将会被逐渐消除，离线机器学习模型的应用，远程生成机器学习模型将成为可能。

**（五）硬件设备发展**

机器学习在从传感器每天收集的大量数据中提取有用信息上发挥着非常重要的作用。在一些应用上，目的是为了分析并理解数据，从而辨清发展趋势（例如，监控、便携式／穿戴式电子设备）。在其他应用中，分析数据的目的是为了能够基于数据快速作出应对（例如，机器人/无人机、自动驾驶汽车、物联网）。对这些应用而言，出于对隐私、安全的考虑，再加上通信带宽的限制，在传感器附近的本地嵌入式处理要比上传到云更好。然而，在传感器端的处理有能耗与成本的限制，还有生产能力与准确率的要求。此外，也需要适应性，以便于传感器适应于不同的应用或环境（例如，在分类器上升级权重与模型）。在许多应用中，机器学习总是涉及到将输入数据转换到更高维度的空间，这伴随着可编程权重、增加数据传输以及最终的能量消耗方面的问题。

时间：从1990 年到2010 年  
速度：非定制CPU 的速度提高了约5000 倍  
优势：因此，现在可以在笔记本电脑上运行小型深度学习模型  
要求：对于计算机视觉或语音识别所使用的典型深度学习模型，所需要的计算能力要比笔记本电脑的计算能力高几个数量级

投资：在20 世纪前十年里，NVIDIA 和AMD 等公司投资数十亿美元来开发快速的大规模并行芯片（图形处理器，GPU），以便为越来越逼真的视频游戏提供图形显示支持。  
推动：这些芯片是廉价的、单一用途的超级计算机，用于在屏幕上实时渲染复杂的3D 场景。这些投资为科学界带来了好处  
历史：

2007 年，NVIDIA 推出了CUDA，作为其GPU 系列的编程接口。少量GPU 开始在各种高度行化的应用中替代大量CPU 集群，并且最早应用于物理建模。深度神经网络主要由许多小矩阵乘法组成，它也是高度并行化的

2011 年前后，一些研究人员开始编写神经网络的CUDA 实现，而Dan Ciresana 和Alex Krizhevskyb 属于第一批人。这样，游戏市场资助了用于下一代人工智能应用的超级计算。有时候，大事件都是从游戏开始的。今天，NVIDIA TITAN X（一款游戏GPU，在2015 年底售价1000 美元）可以实现单精度6.6 TFLOPS 的峰值，即每秒进行6.6 万亿次float32 运算。这比一台现代笔记本电脑的速度要快约350 倍。使用一块TITAN X 显卡，只需几天就可以训练出几年前赢得ILSVRC 竞赛的ImageNet 模型。与此同时，大公司还在包含数百个GPU 的集群上训练深度学习模型，这种类型的GPU 是专门针对深度学习的需求开发的，比如NVIDIA Tesla K80。如果没有现代GPU， 这种集群的超级计算能力是不可能实现的

此外，深度学习行业已经开始超越GPU，开始投资于日益专业化的高效芯片来进行深度学习。

2016 年，Google 在其年度I/O 大会上展示了张量处理器（TPU）项目，它是一种新的芯片设计，其开发目的完全是为了运行深度神经网络。据报道，它的速度比最好的GPU 还要快10 倍， 而且能效更高

二、数据处理框架

**（一）数据获取**

数据获取的手段：数据仓库：数据仓库（DW），数据集市（DM）。检测与抓取，填写，埋点，日志，计算（统计报表某些项目）。

**（二）探索性数据分析与可视化**

单因子与对比分析：集中趋势:均值（分布均匀连续值的趋势）、中位数（对比异常值大小）、离中趋势（标准差，方差）、数据分布（偏态与峰态，正态分布与三大分布）。抽样理论：抽样误差、抽样精度。数据分类：定类（类别）、定序（顺序）、定距（间隔）、定比（比率）。单属性分析:异常值分析：离散异常值、连续异常值、常识异常值。对比分析：绝对数与相对数、时间、空间、理论维度比较。结构分析：考量各组成部分的分布与规律。分布分析：数据分布频率的显式分析。多因子与对比分析：假设检验与方差检验、相关系数：皮尔逊（pearson）与斯皮尔曼（Spearman）。回归：线性回归:关键系数：决定系数、残差不相关（DW检验）。PCA与奇异值分解。

**（三）特征预处理**

特征工程：数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。

特征使用：数据选择、可用性。

特征获取：特征来源、特征存储。

特征处理：数据清洗：数据样本抽样、异常值（空值）处理。

特征选择：剔除与标注不相关的特征、包裹思想遍历特征子集（RFE算法）、嵌入思想（有风险）。

特征变换：对指化、离散化、数据平滑、归一化（标准化）、数值化、正规化。

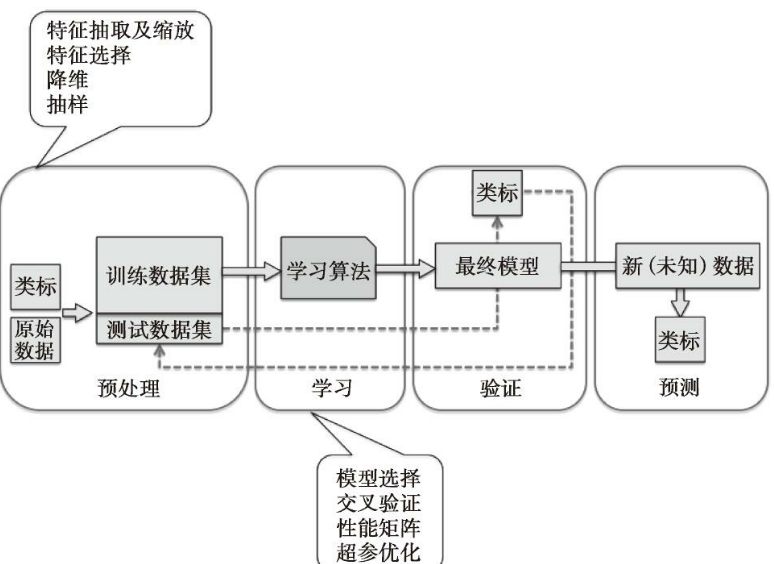
特征降维：PCA,奇异值分解等线性降维、LDA降维，

特征衍生：如加减乘除、求导与高阶求导、人工归纳。特征监控：现有特征、新特征。

**（四）挖掘建模**

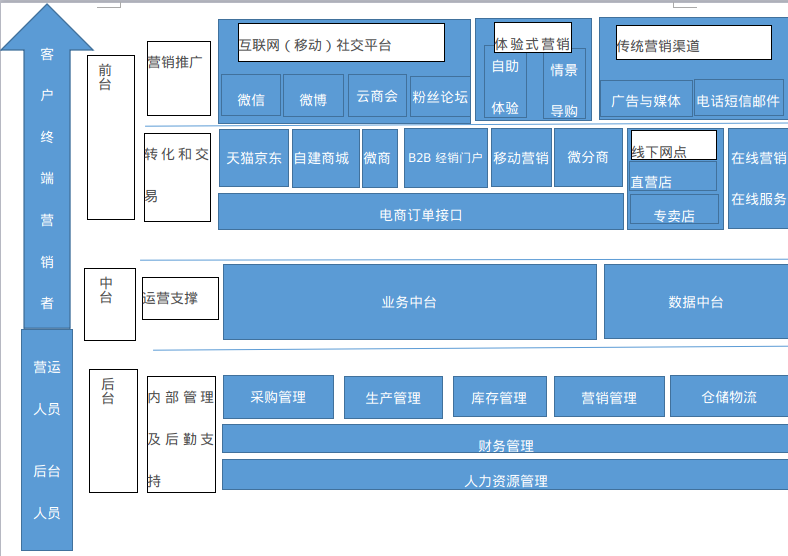
分类方法：KNN,朴素贝叶斯，决策树，支持向量机（SVM），主成份分析（PCA），奇异值分解（SVD）。复合分析：交叉分析、分组与钻取、相关分析、因子分析、聚类分析、回归分析、交叉分析，聚类：K-means、DBSCAN、图分裂，关联：关联规则、序列规则。集成方法:树模型，罗吉斯特映射，人工神经网络。名词解释：学习：通过接收到的数据，归纳提取相同与不同。机器学习：计算机以数据为基础进行归纳总结。模型:数据解释现象的系统。机器学习:监督学习：分类，回归，非监督学习：聚类，关联，半监督学习。数据集：训练集，测试集，验证集。。泛化：对未知数据的预测能力。K-fold交叉验证:将数据集分成k份，每份轮流作一遍测试集，其他作训练集。

**（五）模型评估**

分类模型评估：二分类，混淆矩阵，关键指标：正确率（Accuracy Rate）、召回率（TPR）、F-measure,Precision、错误接收率（FPR）、错误拒绝率（FRR），多元混淆矩阵,ROC与AUC，增益图与KS图。回归模型评估:关键指标:MAE(Mean Absolute Error)、MSE(Mean Square Error)、RMSE(Root MSE)、r2\_score(决定系数)。聚类模型评估:RMS(Root Mean Square),轮廓系数。关联模型评估：支持度，置信度，提升度。

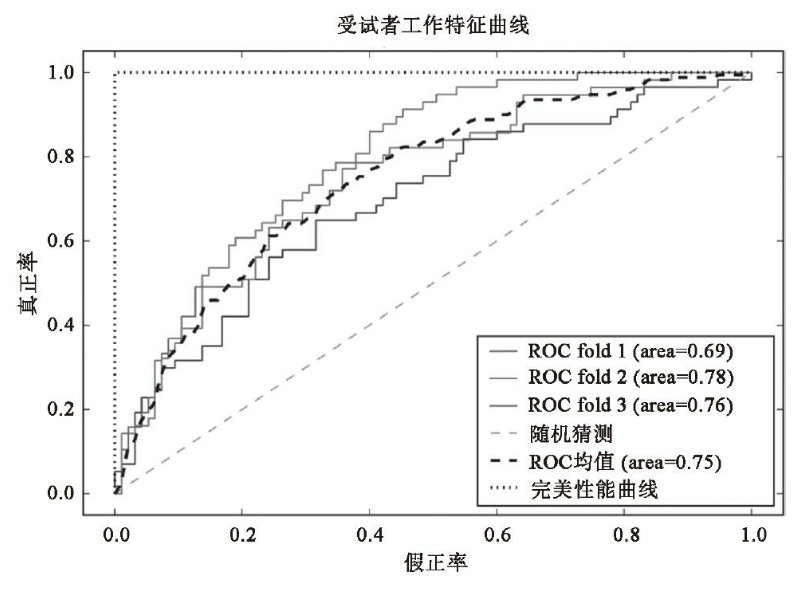
# 三、现代信息技术在会计中的应用构想

计算机本身性能的读存取速度限制了数据处理的速度，古代如果一头牛耕地慢，我们便找来俩头牛来耕种，从来没有见过用把牛喂肥的方法来加快牛耕地的速度，由此分布式读取概念应用而生，既通过多台计算机同时读取数据以提高速度，由此而产生的最著名的分布式数据处理框架hadoop,而我所构想的数据仓库便是以hadoop为基础分布式存储数据。然后利用python和spark进行数据分析。构建数据仓库是非常必要的，因为它是数据分析的基础，公司各个维度的信息通过各种方式存储到数据仓库，构成数据分析的基础。而且本人曾经也想要开发出一个数据处理算法集成的可视化处理软件，其采用了sklearn,scipy,keras,pandas,numpy等库，使用pyqt框架，由于做的时候是读取数据一百多万条数据做的实验没有问题，但是做数据处理的时候为了赶速度用2万条的做测试没有问题，结果到了2019年中国研究生数学建模大赛的时候，我用其做几十列几十万行维度的数据就特别慢了（当然也许是出了点错误），因此后来我才改变思维，那就是一定要做数据分析的基础，数据仓库。随着5G通讯技术的应用落地，云计算技术的成熟打造云上数据仓库已经成为现实，并且可以大幅度降低试错成本。下图为经典数据中台图公司各维度的数据会汇集到数据中台（数据仓库）



举个非常简单的例子，以银行信用卡反诈骗监测为例子，首先数据分布肯定是不均匀的，因为诈骗的人样本一定数量非常少，并且我们要预测的就是诈骗犯,因此需要对数据进行处理。处理方法有俩种上采样和过采样俩种，下采样让多的一个样本和少的一样少，过采样让少的一个样本生成和多的一样多。但是由于我们是为了预测诈骗的样本，所以采用过采样。如果不采用采样方法处理数据我们预测的准确率会非常高，因为预判不出诈骗样本。但是那样高的准确率却是没有意义，因为我们建立模型的目的就是为了

找出诈骗样本。如果采用下采样的话样本结构虽然平衡了但是样本数量却大幅度的减少了，无数的实践案例证明样本数量越多预测结果越好，在许许多多的模型建立前为了提高模型的预测能力大部分时候我们都需要在原有样本基础上随机生成样本。在这里也简单说一下一个模型评判的标准。我们把预测数据分为正类和负类。由此会有：true positives(TP正类判别为正类)， false positives(FP负类判别为正类)，false negatives(FN正类判定为负类)，true negatives(TN负类判定为负类)四种预判结果，召回率Recall=TP/(TP+FN),准确率accuracy=(TN+TP)/(TN+TP+FN+FP)。采用下采样没有影响Recall,影响了精度，采用过采样虽然recall值低了一点，但是整体比下采样好，因为采用的数据多。下图为与此相关常用的ROC曲线图。



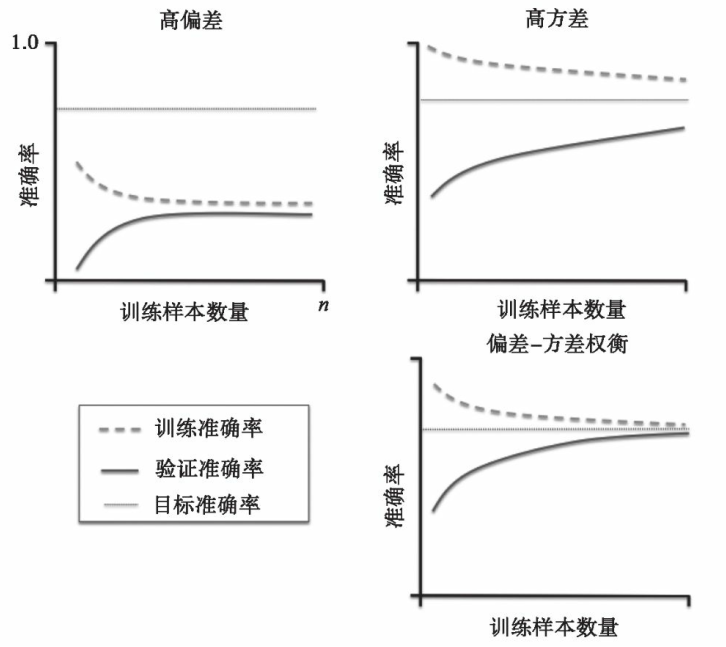
以上废话说了那么多其实也就是数据建模前部曲特征工程的一小部分，事实上我们站在无数巨人的肩膀上建立模型本身是非常简单的，关键的问题是各种问题的思考，数据的预处理简称特征工程，模型评估，发现问题解决问题并前提高模型相关评估指标。

下面在来阐述一下数据建模中的常见问题以及解决方案，一个模型如果参数过多，过于复杂就会造成过拟合，而如果过于简单则又会导致欠拟合。由于维度灾难（curse of dimensionality）的原因，使得KNN算法易于过拟合。维度灾难是指这样一种现象：对于一个样本数量大小稳定的训练数据集，随着其特征数量的增加，样本中有具体值的特征数量变得极其稀疏（大多数特征的取值为空）。直观地说，可以认为即使是最近的邻居，它们在高维空间中的实际距离也是非常远的，因此很难给出一个合适的类标判定。像逻辑斯谛回归类的算法可以通过正则化使其避免过拟合，但是正则化方法并不适用于诸如决策树和KNN等算法，可以使用特征选择和降维等技术来帮助其避免维度灾难。造成过拟合的模型也可以称为模型具有较高的方差，而常用的降低泛化误差的方案有：

1. 收集更多的训练数据
2. 通过正则化引入罚项
3. 选择一个参数相对较少的简单模型
4. 降低数据的维度

LDA的基本概念与PCA非常相似PCA试图在数据集中找到方差最大的正交的主成分分量的轴，而LDA的目标是发现可以最优化分类的特征子空间。LDA与PCA都是可用于降低数据集维度的线性转换技巧。其中PCA是无监督算法，LDA是监督算法。因此我们可以这样客观的认为：与PCA相比，LDA是一种更优越的用于分类的特征提取技术。但是在图像识别任务中的某些情况下，如每个类别中只有少量样本，使用PCA作为预处理工具的分类结果更好。

通过随机森林判定特征的重要性，随机森林有个重要的特性值得一提。例如：如果俩个或者更多个特征是高度相关的，一个特征的相关特征未被完全包含进来，那么此特征可能会得到一个较高的评分。

通过将模型的训练及准确性验证看作是训练数据集大小的函数，并绘制其图像，我们可以很容易看出模型是面临高方差还是高偏差的问题。

左上方图像显示的是一个高偏差模型。此模型的训练准确率和交叉验证准确率都很低，这表明模型未能很好地拟合数据。解决此问题的常用方法是增加模型中参数的数量，例如，收集或构建额外特征，或者降低类似于SVM和逻辑斯谛回归器等模型的正则化程度。右上方图像中的模型面临高方差的问题，表明训练准确度与交叉验证准确度之间有很大差距，针对此类过拟合问题，我们可以收集更多的训练数据或者降低模型的复杂度，如增加正则化的参数，对于不适合正则化的模型，也可以通过特征选择，特征提取来降低特征的数量。需要注意：收集更多的训练数据可以降低模型过拟合的概率。不过该方法不适用于所有问题，例如：训练数据中噪声数据较多，或者模型本身已经接近最优。

硬件设备的不断发展为机器学习提供了养分，硬件设备促进信息收集和机器学习算法的运行，5g网络促进信息传递，云计算为企业实现信息化提供廉价方案。在现在很多中小企业都已经不再自己购买服务器设备，而是直接租用网络服务器，这极大的降低了固定资产的购置费用，设备维护费用，系统维护费用，以及试错成本。而程序开发为了节约成本开发微信小程序，而不是开发Android或者ios应用。

###### 参考文献

1.李航，《统计学习方法》(第二版)，2019年

2.周志华，《机器学习》，2018年

3.Tom White著 王海 华东 刘喻 吕粤海 译，《hadoop权威指南》，2017年

4.Tomasz Drabas Denny Lee著 栾云杰 陈瑶 刘旭斌 译，《PySpark权威指南》，2017年

5.[安德烈亚斯·威蒂格(Andreas Wittig) 迈克尔·威蒂格(Michael Wittig)](https://www.amazon.cn/s/ref=dp_byline_sr_ebooks_1?ie=UTF8&field-author=%E5%AE%89%E5%BE%B7%E7%83%88%E4%BA%9A%E6%96%AF%C2%B7%E5%A8%81%E8%92%82%E6%A0%BC(Andreas+Wittig)+%E8%BF%88%E5%85%8B%E5%B0%94%C2%B7%E5%A8%81%E8%92%82%E6%A0%BC(Michael+Wittig)&text=%E5%AE%89%E5%BE%B7%E7%83%88%E4%BA%9A%E6%96%AF%C2%B7%E5%A8%81%E8%92%82%E6%A0%BC(Andreas+Wittig)+%E8%BF%88%E5%85%8B%E5%B0%94%C2%B7%E5%A8%81%E8%92%82%E6%A0%BC(Michael+Wittig)&sort=relevancerank&search-alias=digital-text)，《AWS云计算实战》，2015年

6.【韩】李在弘 著 武传海 译，《Docker基础与实战》， 2016年

7.Jeff Nickoloff 著 胡震 杨润青 黄帅 译，《Docker实战》，2017年

8.李嘉璇 著，《TensorFlow技术解析与实战》，2017年

9.Peter Harrington 著 李锐 李鹏 曲亚东 王斌 译，《机器学习实战》，2013年

10.Sebastian Raschka 著 高明 徐莹 陶虎成 译，《Python机器学习》，2017年