中国石油大学（华东）
CHINA UNIVERSITY OF PETROLEUM

《计算科学导论》课程总结报告

学生姓名：张宇昊

学　　号：1907010211

专业班级：计科1902

学　　院：计算机科学与技术学院

课程认识 30%	问题思考 30%	格式规范 20%	IT工具 20%	Latex附加 10%	总分	评阅教师

2020年 1月 1日

1 引言

随着社会的进步与科技的发展，计算机成为了人们不可或缺之物，由此诞生了许多学科与行业，而计算机科学便是其中之一。唯有学好计算机科学与技术，才能在这个时代站稳脚跟，不被时代所抛弃。

2 对计算科学导论这门课程的认识、体会

计算科学导论这门课程教导学生如何认识计算机科学与技术，如何学习计算机科学与技术的问题。其中，通过双人分组演讲，我对决策树有了一定的了解；同时，其他小组的演讲，令我对Java、超级计算机、方舟编译器等有了初步的认识。

2.1 计算机科学的基本概念和基本知识

基本概念：狭义的计算科学指称的就是计算机科学与技术，其研究内容涵盖了对计算问题的一般研究。而广义的计算科学包含的内容要广得多，它不仅涵盖了计算机科学与技术的研究范畴计算机科学与技术的研究范畴，而且包含了更多的内涵。（摘自《计算科学导论》（第三版）[5]）

基本知识：计算机科学与技术包括计算机科学、计算机工程、软件工程、信息工程等领域，包含了数字逻辑与集成电路、存储式计算机的基本结构与工作原理、机器指令与汇编语言、计算机网络与通信等方面的知识。

2.2 计算科学发展的现实意义

1.计算机科学与技术能够改善人们的生活

计算机的科学技术发展到今天，计算机设备的价格已经变得十分“亲民”，性能运行更加优良的同时，操作方式也更为简单，外观上也做出了较大的改变。现在计算机已经不再是人们口中的稀罕物件，而是处处可以见到的办公室中的普通办公工具，家庭中的娱乐设施。但是个方面的“亲民”改变没有让人们对计算机失去兴趣，反而更加成为了人们生活中不可或缺的部分，人们的生活、学习、工作许多方面都需要计算机作为辅助，来处理一些事情。计算机的重要性正在被更多的人所意识发现。计算机技术发展所带来的网络技术已经作用于改变人们的生活方式当中。过去我们要购买商品，需要出门、乘车或步行，在商店内进行挑选，然后在原路返回家中，这个过程虽然我们习以为常，却需要时间和空闲的支持。现在我们只需要坐到家中打开手机或电脑，轻点屏幕，不到一个小时新鲜的水果、蔬菜就能配送上门。另外网络中的信息也衍生出了一种新的媒体方式，自媒体。他是来自独立个人的发声，更能体现居民的个人意识，同时大量的信息涌来，也是人们在淹没时增强了分辨信息真实性的能力。更好的享受健康的生活，是计算机技术发展的现实意义之一。

2.计算机科学与技术帮助教育教学

计算机科学技术进入课堂中，为教育领域带来了新的助力。教育的改革是当前教育主旋律，学校与相关教育部门都期待着教育能够向着更加多元化、更加素质化的方向改变。所以校园网络的构建、微课的课堂应用、多媒体技术的教学辅助、网络教育资源的多方共享，所有这

些都是计算机的技术所带来的成果。对于日常学习中信息资料的收集、知识的呈现理解、课后的强化回忆这些也可以由计算机来辅助完成。

2.3 如何学习计算科学与成长

计算机的应用已经深入到生活的方方面面，计算机的发展前景一片光明。随着社会和科技的发展，计算机必将发挥着更为重要的作用。在计算机解决问题的过程中算法是核心。算法是对人解决问题的具体的描述，同时算法是计算机编程的前提，只有有了算法才能够通过编程完成由自然语言转换为机器能够明白的机器语言，才能使计算机完成任务。在计算机解决问题的过程中，计算机是工具。在针对一个生活之中的实际的问题时，需要的是对问题进行算法的描述，在描述中往往要用到多方面的知识，因此，计算机学科是一个涉及到多门学科的学科，不仅如此，还用到了很多生活中的常识。所以，在生活和学习中要掌握名方面的知识，注重专业知识的学习，掌握计算机的原理知识，并且同时不忘进行实践。

3 决策树

结合学习的计算科学知识，对分组演讲涉及的问题作进一步的思考。

3.1 什么是决策树

决策树是一种基本的分类与回归方法。这里我们主要讨论用于分类的决策树，它可以用于机器学习中的决策树训练，使用决策树作为预测模型来预测样本的类标。

定义：分类决策树模型是一种描述对实例进行分类的树形结构。决策树由结点和有向边组成。结点有两种类型：内部结点和叶结点。内部结点表示一个特征和属性，叶结点表示一个类。

可以将决策树看成一个if-then规则的集合。过程：路径上内部结点的特征对应着规则的条件，而叶结点的类对应着规则的结论。

决策树(Decision Tree）是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。

决策树的优点 ：直观，便于理解，小规模数据集有效 。

决策树的缺点 ：处理连续变量不好（例如房价的变化），需要离散化处理；类别较多时，错误增加的比较快；数据规模较大时，计算时间将大大增加，只适合小规模数据集使用。

3.2 进一步的思考

1.过度拟合的定义与避免的方法

百度中关于过度拟合的标准定义：给定一个假设空间H，一个假设h属于H，如果存在其他的假设h’属于H,使得在训练样例上h的错误率比h’小，但在整个实例分布上h’比h的错误率小，那么就说假设h过度拟合训练数据。

有几种途径用来避免决策树学习中的过度拟合。它们可被分为两类：

(1). 及早停止增长树法，在ID3算法完美分类训练数据之前停止增长树；

(2).后修剪法（post-prune），即允许树过度拟合数据，然后对这个树后修剪。

尽管第一种方法可能看起来更直接，但是对过度拟合的树进行后修剪的第二种方法被证明在实践中更成功。这是因为在第一种方法中精确地估计何时停止增长树很困难。无论是通过及早停止还是后修剪来得到正确大小的树，一个关键的问题是使用什么样的准则来确定最终正确树的大小。解决这个问题的方法包括：

(1). 使用与训练样例截然不同的一套分离的样例，来评估通过后修剪方法从树上修剪结点的效用。

(2). 使用所有可用数据进行训练，但进行统计测试来估计扩展（或修剪）一个特定的结点是否有可能改善在训练集合外的实例上的性能。例如，Quinlan （1986）使用一种卡方（chi-square）测试来估计进一步扩展结点是否能改善在整个实例分布上的性能，还是仅仅改善了在当前的训练数据上的性能。

(3). 使用一个明确的标准来衡量训练样例和决策树编码的复杂度，当这个编码的长度最小时停止增长树。这个方法基于一种启发式规则，被称为最小描述长度（Minimum Description Length）的准则。Quinlan & Rivest（1989）和Mehta et al.（1995）也讨论了这种方法。[2]

2.面向大数据分析的决策树

2.1 特征值优化算法

特征值优化算法是指在原有的集合中将数据重新分类，然后形成一个数据子集，对数据子集进行处理分析。特征值优化算法原理较为简单，并且在实践中应用较为简便。利用特征选择值进行算法计算主要可以分为两类，一种是筛选器，一种是封装器。筛选器是指集合内部信息衡量，然后独立于分类算法，这是一个预处理过程。通过相关系数标本进行评价，以达到数据处理的目的。

2.2 集中优化算法

集中优化算法适用于处理数据集合等较为庞大的计算模式，对其内存进行计算过程中没有方法将全部数据内容一次性处理完毕，因此许多数据需要暂时存放在存储器之中。由于决速算法自身的读写操作，因此读写速度比较缓慢，比较适合对这种决策树算法采取优化措施。减少其读写操作的程序成为了决策树算法进行优化的主要方向。在这其中 SICU 就是一种主要的优化算法，这种优化算法通过使用广度排序以及优先原则来达到减少存储器内部读写出生的目的，并且极大提高拳速算法的整体效率，除此之外还有 boat 算法的优化。

2.3 分布式的计算方法

分布式计算方法对其子集进行了扩展，因此在数据处理能力上达到了空前的提高，他能够有效加快数据读取数据的整体能力，并且提高运行的整体速度，因此分布式算法开发比较早。此后谷歌开发了相应的可扩展式的计算机框架这个计算机框架以控制器作为其整体的核心，然后对决策树进行调控调控的主要目的是利用大数据模型来进行整体的训练。同时控制器能够有效接入计算机群中，在学习决策树模型中集成方法也可以解决大数据分布式的问题。

2.4 面向流数据的整体优化算法

流数据整体优化算法可以作为大数据的源头，同时对于叶子阶段相关的统计信息能够有效进行处理，用于代替中间的决策节点，形成新的决策树。在数据整体路以后实现节点分类处理。它能够有效实现统计信息的更新。面向流数据的整体优化算法使得时间成本得到优化，但

是其自身的缺点也很明显，缺乏连续处理素质的能力，同时还可能出现数据的漂流情况。最终的情况会导致大数据信息处理数据准确度有所降低。但是随着现代研究的深入，面向流数据的整体优化算法能够有效支持数值属性的优化处理，因此预测的整体准确性得到了充分的提高，在大数据分析和处理中得到了广泛的应用。[3]

3.随机森林

随机森林主要包括4个部分：随机选择样本；随机选择特征；构建决策树；随机森林投票分类。

3.1.随机选择样本

给定一个训练样本集，数量为N，我们使用有放回采样到N个样本，构成一个新的训练集。注意这里是有放回的采样，所以会采样到重复的样本。详细来说，就是采样N次，每次采样一个，放回，继续采样。即得到了N个样本。

然后我们把这个样本集作为训练集，进入下面的一步。

3.2. 随机选择特征

在构建决策树的时候，我们前面已经讲过如何在一个节点上，计算所有特征的InformationGain（ID3）或者 Gain Ratio（C4.5），然后选择一个最大增益的特征作为划分下一个子节点的走向。

但是，在随机森林中，我们不计算所有特征的增益，而是从总量为M的特征向量中，随机选择m个特征，其中m可以等于sqrt(M)，然后计算m个特征的增益，选择最优特征（属性）。注意，这里的随机选择特征是无放回的选择!

所以，随机森林中包含两个随机的过程：随机选择样本，随机选择特征。

3.3. 构建决策树

有了上面随机产生的样本集，我们就可以使用一般决策树的构建方法，得到一棵分类（或者预测）的决策树。需要注意的是，在计算节点最优分类特征的时候，我们要使用上面的随机选择特征方法。而选择特征的标准可以是我们常见的Information Gain（ID3）或者 Gain Ratio（C4.5）。

3.4. 随机森林投票分类

通过上面的三步走，我们可以得到一棵决策树，我们可以重复这样的过程H次，就得到了H棵决策树。然后来了一个测试样本，我们就可以用每一棵决策树都对它分类一遍，得到了H个分类结果。这时，我们可以使用简单的投票机制，或者该测试样本的最终分类结果。

3.5. 优缺点分析

优点：

它能够处理很高维度（feature很多）的数据，并且不用做特征选择；由于随机选择样本导致的每次学习决策树使用不同训练集，所以可以一定程度上避免过拟合；

缺点：

随机森林已经被证明在某些噪音较大的分类或回归问题上会过拟合；对于有不同级别的属性的数据，级别划分较多的属性会对随机森林产生更大的影响，所以随机森林在这种数据上产出的属性权值是不可信的。[1]

4.决策树是否必须单独使用

答案显然是否定的，决策树可以与其它算法一同使用。例如《一种基于决策树和遗传算法

——BP神经网络的组合预测模型》[4]所写的利用遗传算法对决策树进行调参。又如《基于决策树和遗传算法的神经网络研究及应用》[6]首先对数据挖掘中受到广泛关注的决策树算法、遗传算法和神经网络算法进行了综述,描述了各算法的具体实现过程及步骤。然后通过分析决策树算法与神经网络算法的特点,将它们进行有效得结合,提出一种基于决策树的神经网络权值初始化算法。该算法利用决策树算法,通过分析各样本数据来确定神经网络的初始权值,与传统的神经网络算法比起来,该方法极大的缩小了神经网络初始权值的随机性,使其更有利于最优神经网络模型的生成。最后将该算法应用到了一个通过分析企业类型、注册资金、盈利比例来判断企业信誉的例子中,并通过Matlab编程来实现。文章的第四部分提出的是一种基于遗传算法的神经网络结构优化算法。该算法将遗传算法与神经网络算法进行了巧妙的结合,它利用遗传算法解决了神经网络算法中比较难的结构优化问题,而反过来又巧妙的运用神经网络算法回避了遗传算法中如何确定衡量函数的问题。同样也将该算法应用到了一个超市满意度的例子中,并运用Matlab编程来具体实现了该基于遗传算法的神经网络结构优化算法。

4 总结

计算机科学技术是當前社会必要必须的技术之一，它的功能作用于人们生活，也在不断提升自身的可应用性。更新与发展是它的主旋律，不断为人们的生活带来改变，为教育、为经济、为社会中的各个领域带来助力，促进其发展。但是发展中不能忽视其存在的安全性、技术不完善性等问题，正确面对才能带来更好的发展，发挥其正向积极的作用。同时，这门课程教给我们许多知识，并让我们体会到了合作的重要性，令我们体会到了学好计算机科学与技术的必要性。

5 附录

- 申请Github账户，给出个人网址和个人网站截图

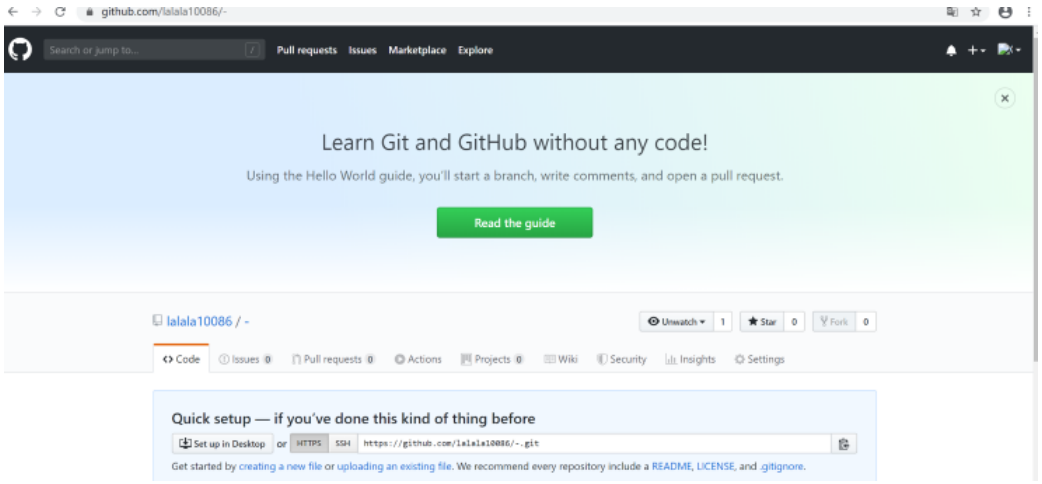

图 1: Github

- 注册观察者、学习强国、哔哩哔哩APP，给出对应的截图

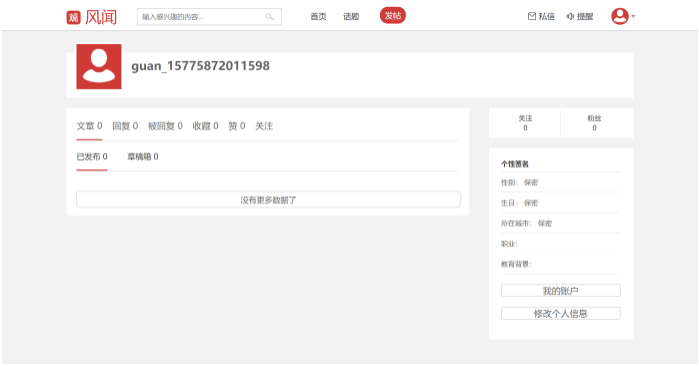

图 2: 观察者

图 3: 学习强国

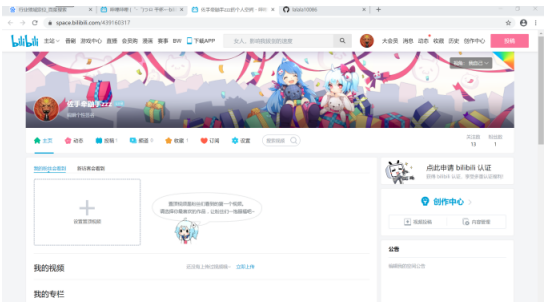

图 4: 哔哩哔哩

- 注册CSDN、博客园账户，给出个人网址和个人网站截图

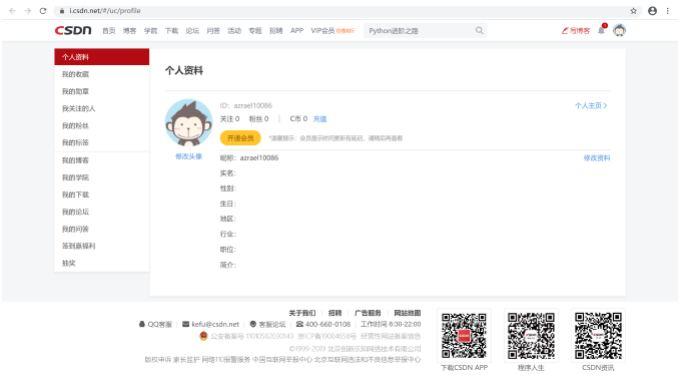

图 5: CSDN

博客园网址：https://www.cnblogs.com/zsqys666/

图 6: 博客园

- 注册小木虫账户，给出个人网址和个人网站截图

网址：http://muchong.com/bbs/space.php?uid=20312905

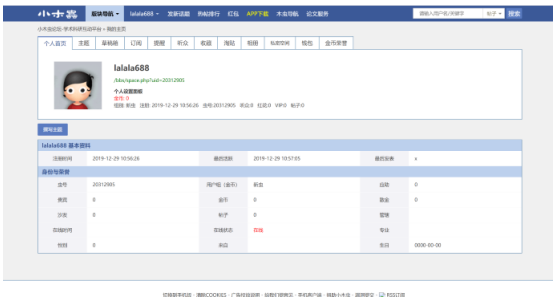

图 7: 小木虫

注意，参考文献至少五篇，其中至少两篇为英文文献，参考文献必须在正文中有引用。

参考文献

[1] 随机森林简介.

[2] Duckie-duckie. 解决决策树的过拟合, 2017.

[3] 杨伟光. 面向大数据分析的决策树算法研究. 电子技术与软件工程, 2018.

[4] 梁栋，张凤琴，陈大武，李小青，王梦非. 一种基于决策树和遗传算法——bp神经网络的组合预测模型. 中国科技论文, 2015.

[5] 赵致琢. 计算科学导论（第三版）. 2006.

[6] 邢远凯. 基于决策树和遗传算法的神经网络研究及应用. 2010.