DOI: 10.3724/SP.J.1224.2019.00254

工 程 管 理

大数据背景下工程投资估算指标编制方法研究

陈志鼎，李 鑫

(三峡大学 水利与环境学院，湖北宜昌 443002)

摘 要: 在现行估算指标逐渐不能满足市场变化需求的情况下，分析了传统估算指标编制方法目前存在的主要问题，提出一种基于大数据技术的估算指标编制方法。通过分析大数据技术在估算指标编制中的应用过程，提出了一种基于数据挖掘 C4.5 算法的估算指标编制方法，综合考虑时间因素对指标的影响，利用敏感性分析对估算指标做一定调整，最后通过算例验证了该方法的实用性，对理论与实践应用均有一定的指导作用。

关键词: 大数据；数据挖掘；C4.5 算法；估算指标编制

中图分类号: TU-9 **文献标识码**: A **文章编号**: 1674-4969(2019)03-0254-11

引言

估算指标是建设项目前期用于编制可行性研究报告、项目建议书的重要依据，属于扩大工程定额的一种，是现代管理科学中的重要内容和基本环节[1]，保证估算指标的市场指导性显得尤为重要。传统估算指标编制方法通常采用具有代表性的工程实例资料，结合现行概预算定额，经过修正、调整后得出综合数据。但由于编制方法受限、耗时较长、工作量大等因素的影响，传统方法编制出的估算指标往往不能很好地反映市场环境变化下的真实情况，同时也耗费了大量的人力物力。

现有的研究主要基于对传统定额编制方法进行改进，如王杰先等[2]从编制定额过程中测定的数据特点出发，构建了数据差异化处理模型，使得处理后的原始数据更加收敛，编制定额的消耗量也更加科学有效。郭琦等[3]则提出一种基于定额消耗量与其影响因素之间的 BP 神经网络算法模型，通过有限的数据样本预测出未知消耗量，省去了部分定额子目难以测定的过程。袁剑波等[4]在高速公路改扩建工程定额的编制过程中应用马氏距离法进行数据处理，使得干扰因素对实测数据的影响明显降低，提高了定额编制的数据精度。朱斌[5]采集了代表性较高的工程样本数据，采用模块化的方法对数据进行了重组，统一了工程规模及费用标准，测算出了典型技术方法下同类工程的投资估算指标。李惠玲等[6]运用灰色模型 GM（1,1）对工程造价指数进行了编制，构建了建筑工程造价指标的预测模型。孟俊娜等[7]将提取的民用建筑的工程特征值作为输入值，以工程造价作为输出值，构建了建筑工程造价的 BP 神经网络估算模型，并通过实例验证其精度确实可行。樊毅等[8]运用模糊数学中的隶属函数，分析了工程项目的实际情况与其特征向量的贴近度，实现了工程估算指标编制。郝晓峰[9]针对初步设计深度不够时，工程量难以准确计算的情况，利用类似工程估算指标进行快速估算编制，并运用估算指标法进行了举例说明。

收稿日期: 2019-03-14; 修回日期: 2019-05-21

作者简介: 陈志鼎（1974-），男，博士，副教授，研究方向为工程项目、造价管理。E-mail：chen_zhiding@163.com

李 鑫（1993-），男，硕士研究生在读，研究方向为工程造价管理。E-mail：164681821@qq.com

以上研究成果大多针对测定数据进行优化处理或改进数据样本的采集方法，但没有从根本上解决定额编制过程周期长、工作量大、工作内容繁复、消耗成本较高等问题。基于上述问题，本文从大数据应用背景出发，提出了一种利用大数据挖掘技术编制估算指标的方法。从大数据技术在估算指标中的应用过程进行分析，重点针对数据利用阶段提出一种数据挖掘 C4.5 算法，用于挖掘满足某些特征的工程实例中实际发生的消耗量，以已完工工程的消耗量来编制估算指标，在考虑时间因素的影响下，用敏感性分析对估算指标做出一定调整，使其能满足时间变化下的投资估算编制需求。

1　大数据技术应用过程

1.1　数据生成

数据生成阶段关心的是数据如何产生。目前数据来源主要有：企业信息系统、web 系统、在线社交网络。以上来源囊括了各种类型的数据，有容易处理的结构化数据，如以 word、excel 形式保存的公司财务报表、合同文本、会议纪要等；还有较难处理的非结构化数据或半结构化数据，如施工现场监测影音资料、网络 html 文件、dwg 格式电子图纸等。

根据《中国建筑施工行业信息化发展报告（2018）》进行的大数据应用情况调研分析[10]，在参与调研的造价站、招标办、建设工程交易中心、建设行政主管部门信息中心等被访对象所在单位收集的数据情况如图 1 所示。

从图 1 可以看出，目前大多数企业数据主要来源是信息系统、电子文档以及纸质文档，以上途径产生的数据大多是简单易利用、数据质量较高的结构化数据。但是根据调研结果，将这些数据利用在工程造价管理信息系统中的占比不足 15%，大量的数据并没有在造价管理中得到深度应用，因而仍需从业人员加强造价管理系统的数据应用开发。

图 1　大数据生成途径及占比

1.2　数据采集

要解决数据的定向挖掘，需要从上层数据的来源考虑，从可以获取数据的途径中收集尽可能多的数据。现阶段工程大数据来源大致分为三个渠道：

1）各大企业的内部数据库，其中包含从可行性研究阶段、设计阶段、施工阶段等整个工程建设全生命周期的所有数据，形式多种多样，主要以文字、报表或图集的形式保存，这类数据目前较难获取，但数据准确性及价值较高；

2）外部市场环境中获取的数据信息，如材料供应商处可获取企业单位的采购清单、机械租赁公司可获取各规格机械的租赁情况、人才市场可获取各单位用工情况等；

3）公开的网络环境，如各大工程类相关网站上的招投标信息、工程实例信息、政府机构公布的开源数据库及行业年报等，这类数据通过开源网络环境较容易获取，但其数据量及有效性有一定的局限性。

1.3　数据存储

目前常用于大数据存储的途径分为集中式存

储和分布式存储，集中式存储即将所有采集的数据集中存储于统一服务器设备，其优点在于方便人为控制、维护方便；分布式存储通常可根据数据类型、体量将其分开存储于不同的独立设备中，其优点在于存储效率较高、数据库易于扩展，可按照需求对数据进行初步分类。

利用分布式存储结合本文数据利用需求（按照项目特征等对数据进行分类挖掘），可引入 SPU（标准化产品单元）概念对工程实例信息等进行特征描述，如描述某住宅项目时可用建设地点、建筑面积、建筑高度、结构型式、装饰标准等代表性特征，其目的在于用最小的特征集合区分不同的工程类别、单位工程、单项工程等，以起到能有效进行数据分层分类存储的效果。

1.4 数据利用

数据利用主要包括数据挖掘分析与落地化应用两部分内涵。数据利用的前提是必须将需求目标与利用技术相结合，如在成本管理中利用数据的目的是成本预测以规避风险、辅助决策等，此时就需要结合成本预测的目标、选择合适的预测算法进行数据挖掘。又如将大数据用于物料管理时，对生产物资部门的出项进项进行追踪，分析出物料使用情况，进行数据累计后反馈问题，对关键物料及影响因素做出分析，以做出最优的物料管理优化方案，对物料进行有效管理，此处需要用到的便是数据挖掘分析算法。

2 基于数据挖掘的估算指标编制方法

2.1 构建数据挖掘 C4.5 算法分类模型

2.1.1 C4.5 算法简介

Quinlan 最初提出了著名的 ID3 算法，此后针对处理连续属性、缺失值情况进行了改进，于 1993 年提出了 C4.5 算法[11]，基于该算法的模型主要用于解决数据挖掘中的分类问题，其基本原理是根据已建立的数据集的特点构造一个用于分类的函数或者模型，使其能够将未知类别的样本映射到给定类别中的某一个。构造分类模型的过程分为三个步骤：

1）将数据集随机地分为训练集和测试集；

2）训练阶段：根据随机分类的训练数据集，通过分析由属性描述的数据元组来构造分类模型，在用于工程项目分类时，可以将建安工程费或人、材、机费用等定义为一个类别，而属性值则可以用该估算指标子目的工作内容、工作类型、适用范围等因素对该类别进行描述；

3）测试阶段：在训练阶段建立分类模型的基础上，利用测试集数据对分类器进行验证，评估数据分类的准确率，如果测试阶段的分类模型能够保证数据分类的准确率，那么可以用该模型进行其他数据组的分类。

2.1.2 计算分类规则参数值并构建分类树

在确定分裂属性构建分类规则树的过程中，最关键的步骤在于确定信息增益率，选择最高信息增益率在叶子节点处作为分裂属性，因此，计算信息增益率在整个分类过程的实现中至关重要。C4.5 算法用于数据挖掘进行数据分类时构造树过程如下。

设 T 为数据集，类别集合为$\{C_1, C_2, \cdots, C_k\}$，选择一个属性 V 把数据集 T 分为若干子集。设 V 有互不重合的 n 个取值$\{V_1, V_2, \cdots, V_n\}$，则 T 被分为 n 个子集 T_1，T_2，…，T_n，这里 T_i 中的所有实例的取值均为 V_i。

令 $\| T \|$ 为数据集 T 的实例个数，$\| T_i \|$ 为 $V=V_i$ 的实例数，$\| C_j \|$ =frep（C_j，T）为 C_j 类的实例数，$\| C_{jv} \|$ 则是 $V=V_i$ 实例中，具有 C_j 类别例子数。

计算类别信息熵：

$$H(C) = \text{info}(T) = -\sum_j \frac{frep(C_j \cdot T)}{\|T\|} \log\left(\frac{frep(C_j \cdot T)}{\|T\|}\right) \quad (1)$$

按照属性 V 把集合 T 分割，分割后的类别条件熵为：

$$H(C|V)=info_v(T)=-\sum_j\frac{\|T_i\|}{\|T\|}\sum_j\frac{\|C_{jv}\|}{\|T_i\|}\log(\frac{\|C_{jv}\|}{\|T_i\|}) \quad (2)$$

计算信息增益：

$$I(C,V)=\text{Gain}(V)\text{info}(T)-info_v(T) \quad (3)$$

属性 V 的信息熵为：

$$H(V)=\text{Split info}(V)=-\sum_{i=1}^{n}\frac{\|T_i\|}{\|T\|}\cdot\log(\frac{\|T_i\|}{\|T\|}) \quad (4)$$

计算信息增益率：

$$\text{Gain ratio}(V)=\frac{I(C,V)}{H(V)}=\frac{Gain(V)}{Split\ info(V)} \quad (5)$$

分别计算每个属性的信息增益率，选取具有最高信息增益率的属性作为分裂属性划分给定的数据集，通过计算各叶子节点的分裂属性可构建分类规则树。

2.1.3 分类树的剪枝优化

由信息增益率确定分裂属性后可以构造分类树，但如果建立的树过于复杂，就会导致生成的分类规则难以理解，甚至可能出现规则冗余的情况。

在工程实例数据的挖掘中存在大量的离散数据，同一分部分项工程，因工作内容、工艺方法、施工条件不同，其人工费、材料费、机械设备费、管理费、利润等属性取值都会不同。分类树构造过程中的算法在计算分裂节点时可能由于属性特征较多，而构造出繁杂的树分支。因此，建立有效的分类树需要考虑树形结构的复杂度，在保证分类准确率的前提下，尽量简化分支，对多余规则进行剪枝。

2.2 估算指标的调整

通过数据挖掘分类算法得到的是已完工工程实例的数据，在进行投资估算编制过程中，由于时间变化，人工价格、材料价格、机械使用费等必然会发生一定的变化，因此需要对时间因素影响下的估算指标做出一定调整。本文通过敏感性分析确定对估算指标影响较大的敏感因素，测算出敏感因素在一定幅度波动导致估算指标的变化，将变化前后数值的比率作为估算指标的调整系数。

3 算例分析

3.1 收集数据集

本文以北京市、上海市两地（2015～2016 年）部分住宅工程为例，收集某些企业内部数据库中的工程实例信息，假设在描述某工程项目时只考虑工程类别、建设地点、建设时间、建筑面积、建筑高度、结构型式、外装饰标准、屋面、内装饰标准、给排水管道、电气管道、给排水工程、电梯等属性特征对其的影响，经过数据汇总后，整理出用上述属性特征描述工程项目的单方造价数据信息，数据样本集见文后附表 1。

3.2 构建分类树

该样本集中属性集有 13 种，即{工程类别、建设地点、建设时间、建筑面积、建筑高度、结构型式、外装饰标准、屋面、内装饰标准、给排水管道、电气管道、给排水工程、电梯}，此处定义单方造价为满足所有属性对应的唯一类别，那么类别集合为{1300，1700，1900，2200}，选取整个数据样本集计算每个属性的信息增益率，选出最高信息增益率的属性作为分裂叶子节点。本文所用数据各属性子集较多，人工计算耗时长，因此不详列计算过程，通过 python 实现 C4.5 算法构造分类树（python 使用版本是基于 windows 运行环境下的 3.7.3 版本，实现过程详见文后附图 1）。

输出分类规则如图 2 所示。

上述导出的分类规则是在分类树建立后进行了树剪枝的结果，算法编程设计中已嵌入预剪枝策略，最后生成的规则有一定的精简，分类规则更加简单易懂，且提高了分类准确率。构建的规则分类树如图 3 所示。

IF 工程类别=多层住宅 AND 建筑高度≤18m AND 结构型式=框架结构THEN单方造价=1300元/㎡

IF 工程类别=多层住宅 AND 建筑高度≤18m AND 结构型式=桩基 AND 外装饰标准=公共部位墙面涂料 THEN 单方造价=1300元/㎡

IF 工程类别=多层住宅 AND 建筑高度≤18m AND 结构型式=PHC桩 THEN 单方造价=1300元/㎡

IF 工程类别=小高层住宅 AND 建筑高度≤45m AND 结构型式=预制方桩THEN 单方造价=1700元/㎡

IF 工程类别=小高层住宅 AND 建筑高度≤45m AND结构型式= 桩基 AND 屋面=挤塑聚苯板保温层 THEN 单方造价=1700元/㎡

IF工程类别=小高层住宅 AND 建筑高度≤45m AND 结构型式=短肢剪力墙THEN 单方造价=1700元/㎡

IF工程类别=中高层住宅 AND 建筑高度≤65m AND 结构型式=预制方桩 THEN 单方造价=1900元/㎡

IF 工程类别=中高层住宅 AND 建筑高度≤65m AND 结构型式=桩基 AND 屋面=憎水珍珠岩砂浆 THEN 单方造价=1900元/㎡

IF 工程类别=高层住宅 AND 建筑高度≤65m AND 结构型式=剪力墙结构THEN 单方造价=1900元/㎡

IF工程类别=高层住宅 AND 建筑高度≤99m AND 结构型式=剪力墙结构 THEN 单方造价=2200元/㎡

IF工程类别=高层住宅 AND 建筑高度≤99m AND 结构型式=桩基 AND 内装饰标准=楼梯环氧树脂地坪 THEN 单方造价=2200元/㎡

IF 工程类别=高层住宅 AND 建筑高度≤99m AND 结构型式=灌注桩 THEN单方造价=2200元/㎡

图 2　输出分类规则

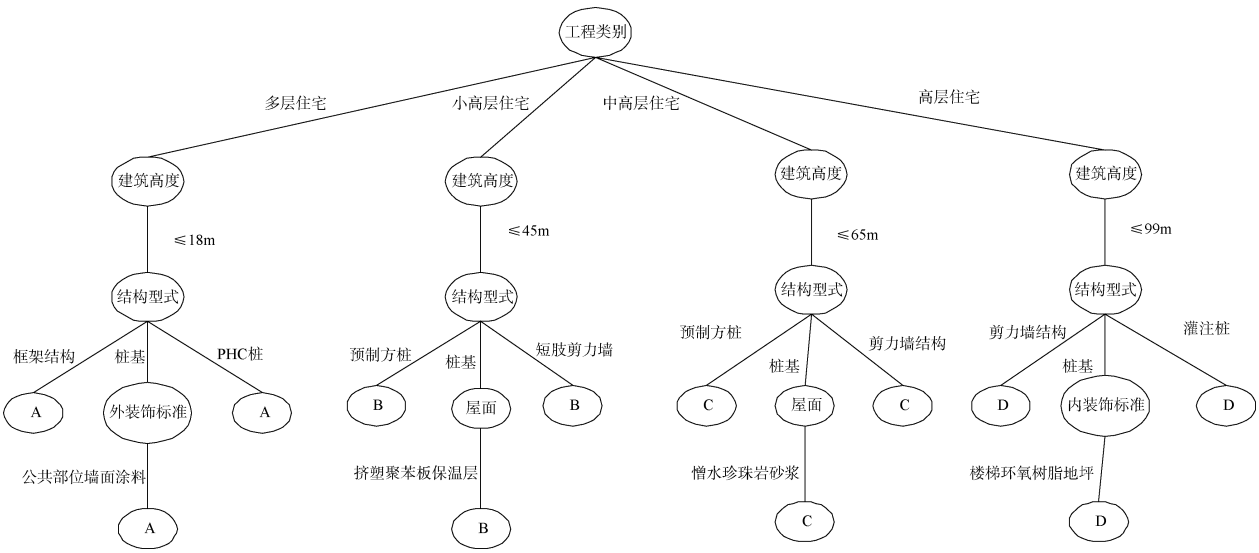

图 3　规则分类树

（注：图中 A、B、C、D 分别对应单方造价 1300 元/m^2、1700 元/m^2、1900 元/m^2、2200 元/m^2）

3.3　测试数据集验证分类规则

通过测试数据集对构造的分类树进行验证，可以将测试数据集中的各描述属性类别代入图 3 的分类树中，得到单方造价，再将包含上述所有描述属性的单位工程估算指标与之对比，得到该分类树在测试数据中的错误率。通常认为，训练样本数与测试样本数保持比例为 1∶3 时，测试结果较为准确，本文因数据样本较多、属性类别较杂，不再列举测试数据验证过程，通过后台同一数据来源渠道随机采集了部分数据样本集用于测试分类树的准确率，测试结果显示错误率为 4.8%。实际应用中，当错误率保持在 10%以下时，可认为该分类树具有较好的实用性。

3.4　估算指标调整系数

在充分考虑时间变化对人、材、机的影响进而导致整个建安工程费发生变化的情况下，可利用单因素敏感性分析法得到影响建安工程费最大的敏感性因素，通过敏感因素变化测算出经济效果指标的变动，从而得到变化前后二者的比率以作为调整系数。分析过程如下。

（1）选择需要分析的不确定因素

在计算建安工程费时，其各项组成部分都可由人工费、材料费、机械使用费计算得来，故本文初步选定人工费、材料费、机械使用费为需要分析的不确定因素，对建安工程费做单因素敏感性分析。经测算，若干工程造价中人、材、机

占比比例约为 2∶7∶1，假定人工费为 20 万、材料费为 70 万、机械使用费为 10 万，工程类别为二类标准，则建筑安装工程费的构成及计算过程如表 1 所示。

表 1　建筑安装工程费的构成

序号	计算方法 / 构成	费用项目		计算公式
一	直接费	直接工程费	人工费	人工费=∑（工日消耗量×日工资单价）
			材料费	材料费=∑（材料消耗量×材料综合单价）
			机械使用费	机械使用费=∑（施工机械台班消耗量×台班单价）
		措施费		根据某地区措施费费率取费标准计取 措施费=（人工费+机械费）×30%
二	间接费	规费		根据某地区规费费率取费标准计取 规费=（人工费+机械费）×10.4%
		管理费		根据某地区管理费费率取费标准计取 规费=（人工费+机械费）×30%
三	利润			利润=（人工费+机械费）×12%
四	税金			税金=（直接费+间接费+利润）×10%

（2）确定分析指标

本文研究对象为估算指标，而单位工程估算指标即建筑安装工程费，将建安工程费作为分析指标，可直观得到时间因素变化导致人工费、材料费、机械使用费在合理范围内变化造成估算指标的变动结果。根据表 1，估算指标的计算公式如式（6）：

单位工程估算指标=建筑安装工程费=直接费+间接费+利润+税金

=[（人工费+材料费+机械使用费）+（人工费+机械使用费）×30%]+（人工费+机械使用费）×（10.4%+30%）+(人工费+机械使用费)×12%+{[（人工费+材料费+机械使用费）+（人工费+机械使用费）×30%]+（人工费+机械使用费）×（10.4%+30%）+(人工费+机械使用费)×12%}×10%

=[（人工费+材料费+机械使用费）+（人工费+机械使用费）×69.4%]×1.1

=2.794×（人工费+机械使用费）+1.1×材料费　　（6）

（3）计算各不确定因素在可能变动范围内发生不同幅度变化导致的经济效果指标的变动结果

由于时间变化，人工综合单价、材料价格、机械使用单价都会发生一定的变化，假设一段时间内定额消耗量不变的前提下，人、材、机的价格在±20%之内以 5%的幅度波动变化，则相应地，经济效果指标即估算指标（建安工程费）如表 2 所示。

（4）确定敏感因素，计算其合理变化范围内引起的指标变化幅度，得到调整系数

通过图 4 敏感性分析可知，在考虑因时间因素导致人工费、材料费、机械使用费发生相同幅度

表 2　不确定因素的变动对建安工程费的影响

不确定因素 \ 变动率 / 建安工程费	–20%	–15%	–10%	–5%	0	5%	10%	15%	20%
人工费	149.64	152.44	155.23	158.03	160.82	163.61	166.41	169.20	172.00
材料费	145.42	149.27	153.12	156.97	160.82	164.67	168.52	172.37	176.22
机械使用费	155.23	156.63	158.03	159.42	160.82	162.22	163.61	165.01	166.41

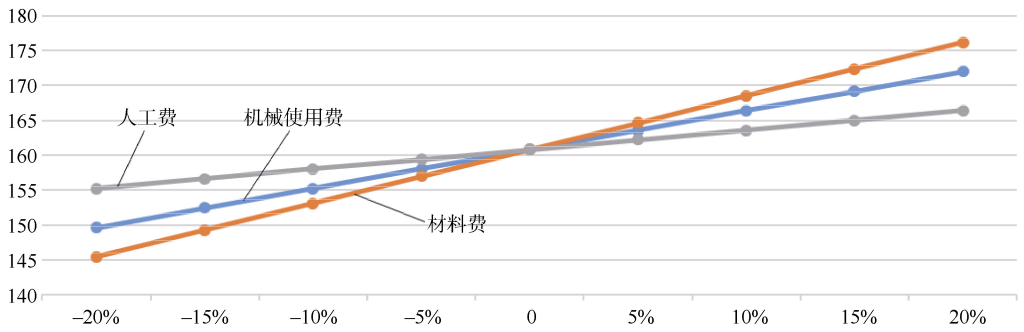

图 4　敏感性分析图

变化的情况下，对建筑安装工程费影响幅度最大的因素是材料价格的变动，因此可确定上述分析的三种可能敏感因素中，材料费为敏感因素。确定敏感因素后，可通过材料费在不同时间的变化对估算指标做一定的调整。在算例中，满足某些特征的住宅工程估算指标为 1300 元/m^2，该指标基于 2015、2016 年的工程实例信息得出，在应用该指标编制某年份相似特征工程的投资估算时，可根据材料价格变化测算出变化后的估算指标，通过时间变化后与变化前的建安工程费比率可得到估算指标的调整系数。

4　结语

（1）针对传统估算指标编制方法存在耗时较长、工作量大、不能快速适应市场变化等问题，结合大数据技术应用过程，提出了一种利用数据挖掘 C4.5 算法对已完工工程实例信息进行数据分类的方法。基于工程实例信息的训练数据集挖掘出满足一定属性特征的分类规则，并通过测试数据集对其进行验证，使得到的分类规则错误率在可以应用的范围内。将分类规则应用于数据分类来编制估算指标，得到的结果更能反映市场真实情况。

（2）综合考虑时间因素对估算指标的影响，通过敏感性分析的方法确定了对估算指标影响最大的因素为材料价格。根据材料价格在未来可能发生的变化，可计算出时间影响后的建安工程费，将变化前后的数值比率作为估算指标的调整系数，调整后的指标具有一定时效性，能有效应对外部市场环境变化带来的影响。

（3）本文对大数据技术在估算指标编制中的应用给出了探索性的方案，但由于目前获取开源数据较困难，文中算例采集的数据体量相对较少，在实际应用中有一定的局限性。因此，开发更多数据采集途径、数据资源整合利用等方面的研究将是下一步探索的方向。

附表 1　2015～2016 年北京、上海部分住宅工程属性特征及数据信息

工程类别	建设地点	建设时间（年）	建筑面积（m^2）	建筑高度（m）	结构形式	外装饰标准	屋面	内装饰标准	给排水管道	电气管道	给排水工程	电梯	单方造价（元/m^2）
多层住宅	上海	2016	4000	≤18	框架结构	普通外墙涂料	挤塑聚苯板保温层	地面防滑地砖	钢塑复合管	塑料管	低水箱坐厕	无	1300
小高层住宅	北京	2016	7000	≤45	预制方桩	外墙涂料	防水砂浆	地面防滑地砖	聚丙烯管	KBG 钢管	洗涤盆	2 台、荷载≤1.11t	1700
多层住宅	上海	2015	3800	≤18	框架结构	普通外墙涂料	挤塑聚苯板保温层	地面防滑地砖	UPVC 雨水管	KBG 钢管	低水箱坐厕	无	1300
中高层住宅	上海	2015	15000	≤65	桩基	胶粉聚苯颗粒保温砂浆	防水卷材	公共部位墙面涂料	UPVC 雨水管	电缆	洗涤盆	3 台、荷载≤1.6t	1900
高层住宅	北京	2016	22480	≤99	灌注桩	局部石料	挤塑聚苯板保温层	楼梯环氧树脂地坪	钢塑复合管	塑料管	低水箱坐厕	荷载≤2.2t	2200
中高层住宅	上海	2016	14500	≤65	桩基	局部石材	防水卷材	楼梯环氧树脂地坪	钢塑复合管	电缆	洗涤盆	3 台、荷载≤1.14t	1900
多层住宅	北京	2015	4000	≤18	桩基	普通外墙涂料	彩色油毡瓦屋面	公共部位墙面涂料	钢塑复合管	塑料管	低水箱坐厕	无	1300
高层住宅	上海	2015	23000	≤99	桩基	外墙涂料	防水砂浆	公共部位墙面涂料	钢塑复合管	KBG 钢管	洗涤盆	荷载≤2.6t	2200
多层住宅	北京	2015	3850	≤18	PHC 桩	胶粉聚苯颗粒保温砂浆	防水卷材	公共部位墙面涂料	钢塑复合管	KBG 钢管	低水箱坐厕	无	1300
小高层住宅	北京	2016	7000	≤45	预制方桩	局部石料	防水砂浆	门厅贴玻化砖	聚丙烯管	塑料管	低水箱坐厕	2 台、荷载≤1.6t	1700
小高层住宅	上海	2015	6200	≤45	短肢剪力墙结构	局部石料	防水卷材	地面防滑地砖	钢塑复合管	塑料管	低水箱坐厕	2 台、荷载≤1.8t	1700
多层住宅	上海	2015	3800	≤18	桩基	胶粉聚苯颗粒保温砂浆	彩色油毡瓦屋面	公共部位墙面涂料	聚丙烯管	KBG 钢管	低水箱坐厕	无	1300
小高层住宅	北京	2015	7000	≤45	预制方桩	外墙涂料	防水砂浆	楼梯环氧树脂地坪	UPVC 雨水管	塑料管	洗涤盆	2 台、荷载≤1.9t	1700
中高层住宅	上海	2016	15000	≤65	剪力墙结构	外墙涂料	挤塑聚苯板保温层	公共部位墙面涂料	钢塑复合管	电缆	洗涤盆	3 台、荷载≤1.11t	1900
小高层住宅	上海	2016	7000	≤45	预制方桩	胶粉聚苯颗粒保温砂浆	PVC 彩色波形瓦屋面	楼梯环氧树脂地坪	钢塑复合管	塑料管	洗涤盆	2 台、荷载≤1.10t	1700
多层住宅	北京	2016	3800	≤18	框架结构	胶粉聚苯颗粒保温砂浆	防水卷材	地面防滑地砖	钢塑复合管	塑料管	低水箱坐厕	无	1300
高层住宅	上海	2015	23000	≤99	灌注桩	外墙涂料	憎水珍珠岩砂浆	门厅贴玻化砖	钢塑复合管	塑料管	低水箱坐厕	荷载≤2.0t	2200

续表

工程类别	建设地点	建设时间（年）	建筑面积（m^2）	建筑高度（m）	结构形式	外装饰标准	屋面	内装饰标准	给排水管道	电气管道	给排水工程	电梯	单方造价（元/m^2）
中高层住宅	北京	2015	15400	≤65	桩基	局部石材	挤塑聚苯板保温层	楼梯环氧树脂地坪	钢塑复合管	电缆	低水箱坐厕	3 台、荷载≤1.13t	1900
小高层住宅	上海	2015	6800	≤45	短肢剪力墙结构	局部石料	防水砂浆	门厅贴玻化砖	聚丙烯管	KBG 钢管	低水箱坐厕	2 台、荷载≤1.12t	1700
高层住宅	上海	2015	22000	≤99	灌注桩	保温砂浆	憎水珍珠岩砂浆	公共部位墙面涂料	UPVC 雨水管	塑料管	低水箱坐厕	荷载≤2.4t	2200
多层住宅	北京	2016	3850	≤18	PHC 桩	胶粉聚苯颗粒保温砂浆	彩色油毡瓦屋面	公共部位墙面涂料	聚丙烯管	KBG 钢管	洗涤盆	无	1300
小高层住宅	上海	2016	6200	≤45	短肢剪力墙结构	胶粉聚苯颗粒保温砂浆	PVC 彩色波形瓦屋面	地面防滑地砖	UPVC 雨水管	KBG 钢管	低水箱坐厕	2 台、荷载≤1.13t	1700
小高层住宅	北京	2015	6800	≤45	桩基	局部石料	防水卷材	门厅贴玻化砖	钢塑复合管	塑料管	洗涤盆	2 台、荷载≤1.7t	1700
中高层住宅	北京	2016	14500	≤65	桩基	局部石材	防水卷材	公共部位墙面涂料	UPVC 雨水管	电缆	洗涤盆	3 台、荷载≤1.10t	1900
高层住宅	上海	2016	22000	≤99	桩基	外墙涂料	憎水珍珠岩砂浆	门厅贴玻化砖	聚丙烯管	KBG 钢管	低水箱坐厕	荷载≤2.5t	2200
中高层住宅	北京	2016	15400	≤65	剪力墙结构	胶粉聚苯颗粒保温砂浆	挤塑聚苯板保温层	公共部位墙面涂料	UPVC 雨水管	KBG 钢管	洗涤盆	3 台、荷载≤1.8t	1900
高层住宅	北京	2016	22480	≤99	灌注桩	胶粉聚苯颗粒保温砂浆	憎水珍珠岩砂浆	公共部位墙面涂料	UPVC 雨水管	电缆	洗涤盆	荷载≤2.3t	2200
中高层住宅	北京	2015	15400	≤65	桩基	胶粉聚苯颗粒保温砂浆	PVC 彩色波形瓦屋面	公共部位墙面涂料	UPVC 雨水管	电缆	洗涤盆	3 台、荷载≤1.9t	1900
多层住宅	北京	2015	3800	≤18	PHC 桩	普通外墙涂料	挤塑聚苯板保温层	公共部位墙面涂料	聚丙烯管	KBG 钢管	低水箱坐厕	无	1300
中高层住宅	上海	2015	15400	≤65	预制方桩	胶粉聚苯颗粒保温砂浆	防水卷材	楼梯环氧树脂地坪	UPVC 雨水管	KBG 钢管	洗涤盆	3 台、荷载≤1.7t	1900
高层住宅	北京	2015	22480	≤99	剪力墙结构	外墙涂料	防水卷材	楼梯环氧树脂地坪	聚丙烯管	KBG 钢管	洗涤盆	荷载≤2.1t	2200
中高层住宅	上海	2016	15000	≤65	剪力墙结构	外墙涂料	挤塑聚苯板保温层	楼梯环氧树脂地坪	聚丙烯管	KBG 钢管	低水箱坐厕	3 台、荷载≤1.12t	1900

数据样本集.py - C:\Users\Lenovo\Desktop\数据样本集.py (3.7.3)

File Edit Format Run Options Window Help

```
from numpy import *
from scipy import *
from math import log
import operator

#计算给定数据的信息熵，
def calcShannonEnt(dataSet):
    numEntries = len(dataSet)
    labelCounts = {}  #类别字典（类别的名称为键，该类别的个数为值）
    for featVec in dataSet:
        currentLabel = featVec[-1]
        if currentLabel not in labelCounts.keys():  #还没添加到字典里的类型
            labelCounts[currentLabel] = 0;
        labelCounts[currentLabel] += 1;
    shannonEnt = 0.0
    for key in labelCounts:  #求出每种类型的熵
        prob = float(labelCounts[key])/numEntries  #每种类型个数占所有的比值
        shannonEnt -= prob * log(prob, 2)
    return shannonEnt;  #返回熵

#按照给定的特征划分数据集
def splitDataSet(dataSet, axis, value):
    retDataSet = []
    for featVec in dataSet:  #按dataSet矩阵中的第axis列的值等于value的分数据集
        if featVec[axis] == value:      #值等于value的，每一行为新的列表（去除第axis个数据）
            reducedFeatVec = featVec[:axis]
            reducedFeatVec.extend(featVec[axis+1:])
            retDataSet.append(reducedFeatVec)
    return retDataSet  #返回分类后的新矩阵

#选择最好的数据集划分方式
def chooseBestFeatureToSplit(dataSet):
    numFeatures = len(dataSet[0])-1  #求属性的个数
    baseEntropy = calcShannonEnt(dataSet)
    bestInfoGain = 0.0; bestFeature = -1
    for i in range(numFeatures):  #求所有属性的信息增益
        featList = [example[i] for example in dataSet]
        uniqueVals = set(featList)  #第i列属性的取值（不同值）数集合
        newEntropy = 0.0
        splitInfo = 0.0;
        for value in uniqueVals:  #求第i列属性每个不同值的熵*他们的概率
            subDataSet = splitDataSet(dataSet, i , value)
            prob = len(subDataSet)/float(len(dataSet))  #求出该值在i列属性中的概率
            newEntropy += prob * calcShannonEnt(subDataSet)  #求i列属性各值对于的熵求和
            splitInfo -= prob * log(prob, 2);
        infoGain = (baseEntropy - newEntropy) / splitInfo;  #求出第i列属性的信息增益率
        print infoGain;
        if(infoGain > bestInfoGain):  #保存信息增益率最大的信息增益率值以及所在的下表（列值i）
            bestInfoGain = infoGain
            bestFeature = i
    return bestFeature
```

数据样本集.py - C:\Users\Lenovo\Desktop\数据样本集.py (3.7.3)

File Edit Format Run Options Window Help

```
#找出出现次数最多的分类名称
def majorityCnt(classList):
    classCount = {}
    for vote in classList:
        if vote not in classCount.keys(): classCount[vote] = 0
        classCount[vote] += 1
    sortedClassCount = sorted(classCount.iteritems(), key = operator.itemgetter(1), reverse=True)
    return sortedClassCount[0][0]

#创建树
def createTree(dataSet, labels):
    classList = [example[-1] for example in dataSet];    #创建需要创建树的训练数据的结果列表（例如最外层的列表是[N, N, Y, Y, Y, N, Y]）
    if classList.count(classList[0]) == len(classList):  #如果所有的训练数据都是属于一个类别，则返回该类别
        return classList[0];
    if (len(dataSet[0]) == 1):  #训练数据只给出类别数据（没给任何属性值数据），返回出现次数最多的分类名称
        return majorityCnt(classList);

    bestFeat = chooseBestFeatureToSplit(dataSet);   #选择信息增益最大的属性进行分（返回值是属性类型列表的下标）
    bestFeatLabel = labels[bestFeat]  #根据下表找属性名称当树的根节点
    myTree = {bestFeatLabel:{}}  #以bestFeatLabel为根节点建一个空树
    del(labels[bestFeat])  #从属性列表中删掉已经被选出来当根节点的属性
    featValues = [example[bestFeat] for example in dataSet]  #找出该属性所有训练数据的值（创建列表）
    uniqueVals = set(featValues)  #求出该属性的所有值得集合（集合的元素不能重复）
    for value in uniqueVals:  #根据该属性的值求树的各个分支
        subLabels = labels[:]
        myTree[bestFeatLabel][value] = createTree(splitDataSet(dataSet, bestFeat, value), subLabels)  #根据各个分支递归创建树
    return myTree  #生成的树

#实用决策树进行分类
def classify(inputTree, featLabels, testVec):
    firstStr = inputTree.keys()[0]
    secondDict = inputTree[firstStr]
    featIndex = featLabels.index(firstStr)
    for key in secondDict.keys():
        if testVec[featIndex] == key:
            if type(secondDict[key]).__name__ == 'dict':
                classLabel = classify(secondDict[key], featLabels, testVec)
            else: classLabel = secondDict[key]
    return classLabel

#读取数据文档中的训练数据（生成二维列表）
def createTrainData():
    lines_set = open('../data/ID3/Dataset.txt').readlines()
    labelLine = lines_set[2];
    labels = labelLine.strip().split()
    lines_set = lines_set[4:11]
    dataSet = [];
    for line in lines_set:
        data = line.split();
        dataSet.append(data);
    return dataSet, labels
```

（接下页）

（接上页）

```
#读取数据文档中的测试数据（生成二维列表）
def createTestData():
    lines_set = open('../data/ID3/Dataset.txt').readlines()
    lines_set = lines_set[15:22]
    dataSet = [];
    for line in lines_set:
        data = line.strip().split();
        dataSet.append(data);
    return dataSet

myDat, labels = createTrainData()
myTree = createTree(myDat,labels)
print myTree
bootList = ['outlook','temperature', 'humidity', 'windy'];
testList = createTestData();
for testData in testList:
    dic = classify(myTree, bootList, testData)
    print dic
```

附图 1　python 实现 C4.5 算法过程

参考文献

[1] 杨万洪. 工程定额管理存在的问题及对策研究[J]. 重庆建筑, 2017, 16(7): 33-34.

[2] 王杰先, 乔　鹏. 工程定额原始数据差异处理模型的建立及应用研究[J]. 价值工程, 2012, 31(27): 88-89.

[3] 郭　琦, 何湘君. 基于 BP 神经网络的水电工程定额编制模型研究[J]. 人民长江, 2016, 47(5): 69-72, 80.

[4] 袁剑波, 毛红日. 基于马氏距离的数据处理方法及其在高速公路改扩建工程定额中的应用[J]. 铁道科学与工程学报, 2018, 15(10): 2715-2720.

[5] 朱　斌. 基于模块化方案的变电站投资估算指标测算方法[J]. 南方能源建设, 2015, 2(S1): 183-186.

[6] 李惠玲, 李晓琴, 刘航天. 建筑工程造价指数的分析与预测[J]. 沈阳建筑大学学报(社会科学版), 2013(1): 74-77.

[7] 孟俊娜, 梁　岩, 房　宁. 基于 BP 神经网络的民用建筑工程造价估算方法研究[J]. 建筑经济, 2015(9): 64-68.

[8] 樊　毅. 基于模糊数学方法的输电线路工程造价估算研究[D]. 华北电力大学, 2012.

[9] 郝晓峰. 运用概算指标法编制单位工程概算[J]. 中小企业管理与科技(下旬刊), 2010(8): 207.

[10] 中国建筑施工行业信息化发展报告（2018）大数据应用与发展[M]. 中国建材工业出版社, 2018.

[11] 于　露, 杨亚宁. C4.5 算法的研究与应用[J]. 云南大学学报(自然科学版), 2010, 32(S2): 136-140.

Compiling Engineering Investment Estimation Indicators Using Big Data

Chen Zhiding, Li Xin

(*College of Hydraulic & Environmental Engineering*, *Three Gorges University*, *Yichang*, *Hubei* 443002, *China*)

Abstract: Presently, estimation indicators used for engineering investment calculations are unable to meet market demands. We analyze the major issues that affect traditional estimation index preparation methods, and we propose a big data-based method for compiling estimation indicators. This new method uses data mining and the C4.5 algorithm. Considering the influence of time on the estimation index, sensitivity analyses were conducted to adjust certain estimation indicators. The effectiveness of the method is verified using a practical example, which has a certain guiding effect on both theory and practice.

Key Words: big data; data mining; C4.5 algorithm; estimation indicators compilation