doi:10.3969/j.issn.1673-9833.2019.05.010

统计分析及决策树算法在高校就业指导中的应用

韩存鸽

（武夷学院 数学与计算机学院，武夷山 福建 354300）

摘 要：对福建某高校二级学院 2016 年毕业生就业信息进行预处理，从统计分析、构建决策树模型两方面开展研究。在 Weka 中采用 C4.5（J48）算法构建了决策树，根据分析及构造的决策树模型，从人才培养方案、奖励制度、在校学生的就业规划提出了相关建议，为就业指导部门和高校领导提供一定的决策帮助。

关键词：统计分析；决策树；C4.5(J48) 算法；Weka

中图分类号：TP181 **文献标志码**：A **文章编号**：1673-9833(2019)05-0057-05

引文格式：韩存鸽 . 统计分析及决策树算法在高校就业指导中的应用 [J]. 湖南工业大学学报，2019，33(5)：57-61.

Application of Statistical Analysis and Decision Tree Algorithm in College Employment Guidance

HAN Cunge

（Mathematics and Computer Science College，Wuyi University，Wuyishan Fujian 354300，China）

Abstract：This paper preprocesses the employment information of the graduates from a secondary college in Fujian Province in 2016, with the research carried out from the statistical analysis and decision tree model construction. The decision tree is to be constructed by using C4.5 (J48) algorithm in Weka. Based on the analysis and construction of the decision tree model, this paper puts forward relevant suggestions from personnel training scheme, reward system and employment planning of students in colleges, thus providing certain decision-making help for employment guidance departments and college leaders.

Keywords：statistical analysis；decision tree；C4.5(J48) algorithm；Weka

近年来，随着高校招生规模逐年扩大、毕业人数逐年增加，高校毕业生的就业形势日益严峻。现如今，各高校都建立了比较完善的学生信息管理系统，积累了大量的历史数据。而现有的数据库虽然可以很好地实现对这些数据的存储、查询和维护等功能，但是无法发现这些数据中存在的关系和规则，缺乏综合分析和辅助决策的能力 [1]。本研究拟从现有的学生信息中，发掘影响大学生就业的主要因素，为就业指导部门和高校领导提供一定的决策帮助，也为在校学生的职业规划提供指导帮助。

1 决策树与 C4.5 算法

1.1 决策树概念及分类过程

国内外对数据挖掘的研究主要集中在分类、聚类、关联规则挖掘、序列模式发现、异常和趋势发现等方面。分类挖掘技术在商业、科研、银行等领域中的成功应用，使其成为了数据挖掘中最活跃、最成熟的研究方向。其中决策树 [2]（decision tree）是

收稿日期：2018-12-03

基金项目：国家自然科学基金资助项目（51272074）

作者简介：韩存鸽（1979-），女，陕西咸阳人，武夷学院教师，硕士，主要研究方向为数据库和数据挖掘，E-mail：gezi0401@163.com

最常见的一种分类方法。在数据挖掘中该技术主要应用于分类和预测 [3]。分类过程如下：通过训练子集建立一棵决策树，如果该树不能对所有对象给出正确的分类，则选择一些其他的训练子集加入原来的训练子集中，重复该过程直到形成正确的决策树，决策树分类过程 [4] 如图 1 所示。

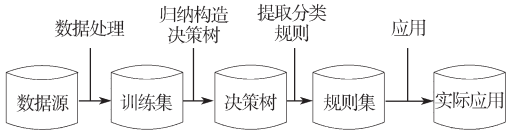

图 1 决策树分类过程

Fig. 1 Decision tree classification process

早期著名的决策树算法是基于信息熵的 ID3 算法 [5]，后来在 ID3 的基础上，又提出了各种决策树方法，如 C4.5、C5.0、ID4、CART、SLIQ 等算法。本研究中使用 C4.5 算法。

1.2 C4.5 算法理论

C4.5[6] 算法是在 ID3 算法的基础上发展的决策树算法，其采用信息增益率作为属性选择的度量依据，在处理连续性属性时，C4.5 算法将这些连续属性“离散化”，解决了 ID3 算法多值偏向性的缺点，有效避免了取值较多的属性容易被选作分裂属性的问题，增强了决策树模型的有效性。该算法准确率较高，曾受到人们广泛的关注和应用。为了研究方便，下面就 C4.5 算法的相关概念 [7] 进行说明。

1）期望信息（也称信息熵）

设数据 S 为类标记元组的训练集。假定类标号属性有 m 个不同值，定义 m 个不同类 T_i（i=1, 2,⋯, m）。设 $|S|$ 和 $|T_i|$ 分别是 S 和 T_i 类中元组的个数，对一个给定的样本分类所需的期望值为

$$Info(S)=-\sum_{i=1}^{m}\frac{|T_i|}{|S|}\log_2\frac{|T_i|}{|S|}。$$

2）信息增益

设 s_{ij} 是子集 s_j 中类 T_i 的样本数，将属性 A 划分成 v 个子集，其信息量为：

$$E(A)=Info_A(S)=\sum_{j=1}^{v}\frac{|T_i|}{|S|}Info(s_{ij})。$$

信息增益为原来的信息需求与新的需求之间的差，即 $Gain(A)$=$Info(S)$−$E(A)$。

3）信息增益率

为了使算法更健壮稳定，在 C4.5 算法中引入信息增益率，则属性 A 的信息增益率计算公式如下：

$$GainRatio=\frac{Gain(A)}{SplitE(A)};$$

$$SpliE(A)=-\sum_{i=1}^{k}\frac{|T_i|}{|S|}\log_2\frac{|T_i|}{|S|}。$$

1.3 C4.5 算法生成决策树过程

假设 S 为训练样例集合，样例的候选属性集用 A 表示，采用 C4.5 算法产生决策树，具体描述如下 [8]：

1）创建根结点 N；2）如果训练集为空，则返回根结点 N，并以失败结点标记；3）如果训练样本集中所有的样本都属于一个类别，则把结点 N 标记为该类别；4）如果没有可选属性，则返回 N 作为叶子结点，直接标记为最常出现的类别；5）计算每个候选属性的信息增益率 $GainRatio$；6）选择信息增益率最大的属性作为划分当前数据集的分裂属性；7）根据分裂属性的取值划分训练集，在每个训练子集上递归地执行 C4.5 算法，得到初步决策树；8）利用更大的训练数据集对决策树进行修剪（优化）。

1.4 C4.5 算法执行流程

C4.5 算法采用后剪枝技术对生成的决策树进行剪枝操作，形成决策树模型，根据建立好的模型，生成一系列 IF−THEN 规则，实现对训练集的分类 [9]。C4.5 算法生成决策树时采用自上而下、分而治之的思想，依次选取每个属性作为节点，自顶向下生成一棵树。其执行流程如图 2 所示。

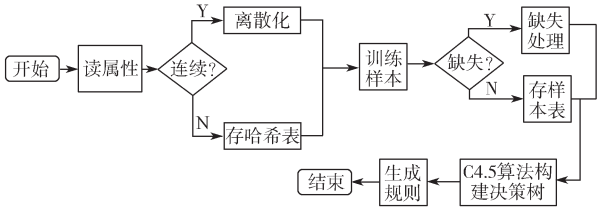

图 2 C4.5 算法执行流程图

Fig. 2 C4.5 algorithm execution flow chart

C4.5 算法的优点是能够产生准确率高、易于理解的分类规则，缺点是该算法在构造树的过程中，需要多次对数据集进行顺序扫描及排序 [10]，从而导致算法低效，同时 C4.5 算法伸缩性不好，只能处理驻留于内存的数据集，因此要求训练集不能过大 [11]。

2 学生就业信息的预处理及统计分析

2.1 数据预处理

高校的各个系统运行中，可能会存在数据缺失、数据不准确、数据不一致等问题，因此需要对源数据进行预处理，数据预处理分为以下步骤：数据集成、数据抽取、数据清洗、数据规范。

1）数据集成

本文的原始数据来源于福建某高校 2016 年毕业

生就业信息，主要涉及学生基本信息表、学生成绩表、四六级成绩表、学生就业管理信息表。最终用到的就业总表是以这 4 个表为基础形成的关系数据库。

在 SQL Server 2008 中通过“身份证号”“学号”把学生基本信息表、学生成绩表、四六级成绩表、学生就业管理信息表 4 个表匹配并链接形成就业总表 Jyzbc，就业分析的星形模式如图 3 所示。

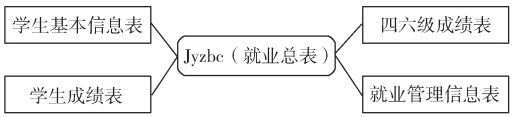

图 3　就业分析的星形模式

Fig. 3　Star pattern of employment analysis

Jyzbc 属性为（学号、姓名、性别、身份证号、民族、专业、政治面貌、社会实践能力、成绩、四六级成绩、就业单位、单位联系人、单位地址、单位性质）。

2）数据抽取

在获取的 Jyzbc 表中，含有过多的属性，由于学号、姓名、民族、身份证号、单位联系人、单位地点等属性与学生就业情况并无太大的关联，挖掘这些属性没有意义，所以将这些属性删除。因为学号、实践能力、专业成绩、英语等级、就业情况这些属性与挖掘结果关系密切，所以予以保留。

3）数据规范

为了方便分析，课题组对专业成绩、英语等级、实践能力、单位性质进行分类。学生专业成绩取学生在校期间所有必修课的平均成绩，分优秀（专业成绩不低于 85 分，excellent）、良好（专业成绩不低于 80 分但低于 85 分，good）、中等（专业成绩不高于 79 分，average）3 个等级。英语等级分优（六级及以上，excellent）、良（四级，good）、一般（四级以下，average）3 个类别，实践能力分 3 个等级：优（在校期间工作过或在外实习经历最少 3 次，excellent）、良（在校期间参加社会实践或实习 2 次，good）、中（在校期间参加社会实践或实习 1 次及以下，average）。单位性质分为企业（QY）、事业单位（SY）、自主创业（CY）、考研升学（KY）4 大类，其中，企业又分为国企（QY1）、外企（QY2）和私企（QY3）。其中专业成绩的数据范化 T-SQL 语句如下：

```
select *,zycj=
  case
    when 专业成绩 <=79 then 'average'
     when 专业成绩 >=80 and 专业成绩 <85 then 'good'
    when 专业成绩 >=85 then 'excellent'
  end
  into jyzb1
  from jyzbc
```

对英语等级、实践能力、单位性质数据进行规范与专业成绩处理类似，最后形成 Jyzbc 表。

4）数据清洗

在数据的日常管理维护中由于各种原因，难免会出现不完整的数据，或者在录入数据的过程中出现某些误差，因此有必要对表中不符合规范的数据予以删除，最后将清理完的信息列出，共有 435 条完整的记录，经过处理后 jyzb.csv 中部分数据如图 4 所示。该数据集中有 5 个属性，其中 Xh 为数据序号，zycj 为专业成绩，yydj 为英语等级，sjnl 为实践能力，Dwxz 为单位性质。

Xh	zycj	yydj	sjnl	Dwxz
1	good	good	excellent	QY1
2	excellent	excellent	good	SY
3	good	good	good	QY2
4	good	good	excellent	QY2
5	excellent	good	good	SY
6	average	good	good	SY
7	good	excellent	good	SY
8	excellent	excellent	good	SY
9	excellent	good	average	QY1
10	average	good	good	SY

图 4　预处理后的部分数据

Fig. 4　Partial data tables after preprocessing

2.2　统计分析在就业指导中的应用

根据预处理的数据按就业单位性质统计，得出的就业情况如图 5 所示。

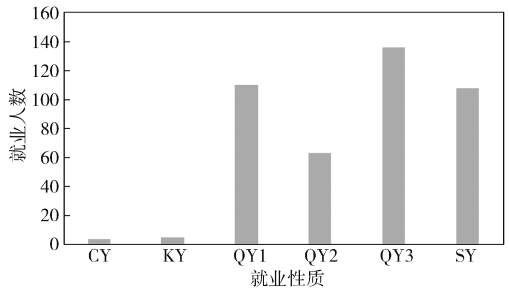

图 5　就业情况统计

Fig. 5　Employment statistics

由图 5 可知，其中进入 QY3（私企）、QY1（国企）、SY（事业单位）这 3 种单位就业的学生人数较多，CY（自主创业）、KY（考研升学）这两类的数据量很少。故这两类数据无法产生规则，在后面的决策树分析中，删除这两类产生的 9 条数据，最后进行决策树分析的有 426 条数据。

根据以上数据，建议学校建立考研奖励制度，鼓励学生继续升学深造；大力宣传大学生自主创业优惠政策，鼓励毕业生自主创业，做好创业指导工作。根据进入 QY2（外企）的学生人数较少这种情况，建议学校修改人才培养方案，建立校企合作制度。学校根据用人单位的人才需求，开设相应的课程，为企业“订单式”定向培养学生，同时邀请企业定期安排

专业技术人员到学习开设讲座，把前沿、先进的专业技术带进课堂。

3 决策树算法在就业指导中的应用

3.1 Weka 挖掘平台介绍决策树

Weka 的全名是怀卡托智能分析环境（Waikato environment for knowledge analysis）[12]，其中汇集了大量机器学习算法和数据预处理工具，包括分类、回归、聚类、关联规则及在新的交互式界面上的可视化。本研究通过对学生就业数据的分析，选择 Weka 作为数据挖掘工具。在 Weka 中使用 J48（C4.5）算法进行测试，找出影响大学生就业的主要因素。

3.2 J48 算法在就业指导中的应用

Weka 中的决策树算法有 BFTree、FT、LADTree、DecisionStump、LMT、NBTree 等 13 种，这里选用经典的决策树算法 J48，在 Weka 平台中将 C4.5 算法命名为 J48 算法，以下均称 J48 算法。启动 Weka 导入 Jyzb.csv 训练数据集，选择 J48（C4.5）算法进行测试，J48 分类器默认参数为“J48 – C0.25 –M2”，其中参数“–C0.25”表示置信因子，“–M2”表示最小实例数。挖掘结果见图 6，总共有 426 个实例参与分类，有 362 个实例被准确分类，64 个实例被错误分类，Correctly Classified Instances 为 84.976 5%，说明模型的准确率较高。Kappa Statistic 为 0.732 9，该值用来评判分类结果与随机分类的差异度，越接近 1 分类结果越好，分类结果说明分类效果较好。

```
=== Summary ===

Correctly Classified Instances         362               84.9765 %
Incorrectly Classified Instances        64               15.0235 %
Kappa statistic                          0.7329
Mean absolute error                      0.1127
Root mean squared error                  0.2386
Relative absolute error                 34.3809 %
Root relative squared error             59.008  %
Total Number of Instances              426

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC A
                 0.773     0          1         0.773     0.872      0.9
                 1         0.116      0.745     1         0.854      0.9
                 1         0.069      0.571     1         0.727      0.9
Weighted Avg.    0.85      0.035      0.899     0.85      0.855      0.9

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
```

图 6 基于 J48 算法的决策分类结果

Fig. 6 Decision classification results based on J48 algorithm

3.3 决策结果分析

在系统中选择 Visualize Tree（可视化树），可以看到生成的决策树，如图 7 所示。

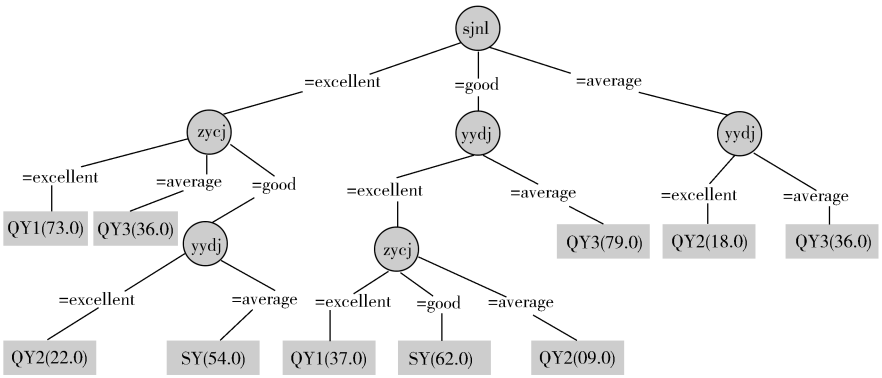

图 7 构建的决策树

Fig. 7 Decision tree to be constructed

“IF–THEN”是一种简单的表示分类规则的有效方法，遍历决策树就可以形成一条从根节点到叶节点的规则，根据决策树模型提取如下规则。

Rule 1：IF sjnl=excellent AND zycj=excellent THEN QY1。

Rule 2：IF sjnl=excellent AND zycj=average THEN QY3。

Rule 3：IF sjnl=excellent AND zycj=good AND yydj=excellent THEN QY2。

Rule 4：IF sjnl=excellent AND zycj=good AND yydj=average THEN SY。

Rule 5：IF sjnl=good AND yydj= excellent AND zycj=excellent THEN QY1。

Rule 6：IF sjnl=good AND yydj= excellent AND zycj=good THEN SY。

Rule7：IF sjnl=good AND yydj= excellent AND zycj=average THEN QY2。

Rule 8：IF sjnl=good AND yydj= average THEN QY3。

Rule 9：IF sjnl=average AND yydj= excellent THEN QY2。

Rule 10：IF sjnl=average AND yydj= average THEN QY3。

每条规则都有前件和后件，如 Rule1 解释如下：如果某位学生 sjnl 和 zycj 都为优，那么该同学很可能去国企工作。从决策树规则可以看出，sjnl（实践

能力）、zycj（专业成绩）、yydj（英语等级）对学生就业的影响较大，实践能力好的学生，就业面比较广，英语成绩好的学生，进入 QY1（国企）、QY2（外企）、SY（事业单位）的机率就高；而专业成绩一般，英语成绩一般的学生，基本就在 QY3（私企）工作，就业面比较窄，就业机会也比较少。这些规则也为低届学生提供就业参考，在校事先做好准备，如果学生毕业后若想去国企，就要侧重专业课程的学习，努力提高专业课的成绩，若想去 QY2（外企）工作，则需要加强英语和实践动手能力两方面的培养；若去 SY（事业单位）上班，就应该加强实践能力和专业课程的学习。相对应学校就业指导部门如果想提高学生的就业率，针对不同的就业单位采取不同的策略。若想提高去 QY1（国企）的就业率，应该加强学生专业能力的培养；要提高去 QY2（外企）的就业率，应该加强学生英语和实践能力的培养；要提高去 SY（事业单位）的就业率就应该加强专业能力和实践能力的培养；否则将导致多数学生到私企就业。

4　结语

关联规则挖掘虽然可以有针对性地挖掘出学生就业需要具备的素质，但是需要慎重选择挖掘对象和制定阈值标准，这样才能保证挖掘结果的有效性。本研究对福建某高校二级学院 2016 年毕业生就业信息进行预处理，从统计分析、构建决策树模型两方面开展工作。

根据统计结果，建议学校建立考研奖励机制，做好大学生自主创业政策的宣传与推广工作，同时修改已有的人才培养方案，建立校企合作制度，以提高大学生就业率。

在 Weka 中采用 C4.5（J48）算法构建大学生就业决策模型，挖掘并分析就业规则，找出影响大学生就业的主要因素，给在校学生的职业规划提供帮助，为就业指导部门和高校领导提供一定的决策帮助。学校在今后的就业指导中，可以根据岗位需求加强相关培训，帮助学生找到合适的工作，实现毕业生更快、更好地就业。

参考文献：

[1] 张娅妮 . 数据挖掘技术在就业指导中的应用研究 [J]. 淮海工学院学报 (自然科学版)，2013，22(2)：32-34.
ZHANG Yani. Application of Data Mining Technology to Career Guidance[J]. Journal of Huaihai Institute of Technology(Natural Science Edition)，2013，22(2)：32-34.

[2] GRANA C，MONTANGERO M，BORGHESANI D. Optimal Decision Trees for Local Image Processing Algorithms[J]. Pattern Recognition Letters，2012，33(16)：2302-2310.

[3] 李　旭 . 五种决策树算法的比较研究 [D]. 大连：大连理工大学，2011.
LI Xu. A Comparative Study on Five Decision Tree Algorithms[D]. Dalian：Dalian University of Technology，2011.

[4] DUNHAM M H. 数据挖掘教程 [M]. 北京：清华大学出版社，2004：188.
DUNHAM M H. Data Mining Course[M]. Beijing：Tsinghua University Press，2004：188.

[5] CHEN J，LUO D L，MU F X. An Improved ID3 Decision Tree Algotithm[C]// 2009 4th International Conference on Computer Science & Education. Nanning：IEEE，2009：127-130.

[6] 乐明明 . 数据挖掘分类算法的研究和应用 [D]. 成都：电子科技大学，2017.
LE Mingming. Research and Application of Data Mining Classification Algorithm[D]. Chengdu：University of Electronic Science and Technology of China，2017.

[7] HAN J W，KAMBER M. 数据挖掘概念与技术 [M]. 北京：机械工业出版社，2006：191-192.
HAN J W，KAMBER M. Concept and Technology of Data Mining[M]. Beijing：China Machine Press，2006：191-192.

[8] 徐　鹏，林　森 . 基于 C4.5 决策树的流量分类方法 [J]. 软件学报，2009，20(10)：2692-2704.
XU Peng，LIN Sen. Internet Traffic Classification Using C4.5 Decision Tree[J]. Journal of Software，2009，20(10)：2692-2704.

[9] 缪连芬 . 改进的 C4.5 算法在大学生情感素质分析中的研究与应用 [D]. 上海：上海师范大学，2018.
MU Lianfen. Research and Application of Improved C4.5 Algorithm in Emotional Quality Analysis of College Students[D]. Shanghai：Shanghai Normal University，2018.

[10] KAPLAN L，BRONSTEIN Y，BARZILAY Y，et al. Canal Expansive Laminoplasty in the Management of Cervical Spondylotic Myelopathy[J]. Israel Medical Association Journal，2006，8(8)：548-552.

[11] HASHIM H，TALAB A A，SATTY A，et al. Data Mining Methodologies to Study Student's Academic Performance Using the C4.5 Algorithm[J]. International Journal on Computational Science & Applications，2015，5(2)：59-68.

[12] 袁梅宇 . 数据挖掘与机器学习 Weka 应用技术与实践 [M]. 北京：清华大学出版社，2016：2.
YUAN Meiyu. Application Technology and Practice of Data Mining and Machine Learning Weka[M]. Beijing：Tsinghua University Press，2016：2.

（责任编辑：申　剑）