Computer Science and Application 计算机科学与应用, 2013, 3, 80-84
http://dx.doi.org/10.12677/csa.2013.31014 Published Online February 2013 (http://www.hanspub.org/journal/csa.html)

The Correlate Analysis of Housing Conditions and the User Related Factors Based on Data Mining

Jian Sun, Yunlong Zhou

College of Communications, University of Electronic Science and Technology of China, Chengdu
Email: yunlongz@163.com

Received: Nov. 17th, 2012; revised: Nov. 29th, 2012; accepted: Dec. 11th, 2012

Abstract: Decision tree is an important means of data mining. It has a wide range of applications in data mining knowledge discovery. The paper uses SQL Server Business Intelligence Development Studio platform. Using decision tree data model, we can draw decision tree. At last, we analyse and predict the results, so that we draw a more ideal conclusion.

Keywords: Data Mining; Decision Tree; SQL; Housing Conditions

基于数据挖掘的住房状况与用户相关因素分析

孙　健，周云龙

电子科技大学通信学院，成都
Email: yunlongz@163.com

收稿日期：2012 年 11 月 17 日；修回日期：2012 年 11 月 29 日；录用日期：2012 年 12 月 11 日

摘　要：决策树是数据挖掘的一种重要手段，在数据挖掘知识发现中有广泛的应用。本文在 SQL Server Business Intelligence Development Studio 平台上，通过决策树模型绘制了决策树并且得出了关于预测项住房状况的影响因子以及影响程度的强弱，最后对数据挖掘结果进行分析与预测且得到了比较理想的预测与结论。

关键词：数据挖掘；决策树；SQL；住房状况

1. 引言

本文是利用 SQL Server 数据挖掘对大规模数据集 MovieClick 进行挖掘，以便从大量繁杂的数据中获取隐含中其中的信息[1,2]。MovieClick 数据库是通过收集客户喜欢的电影的相关内容以及客户自身数据的一个数据集，如 Num bedrooms、Num cars、Marry Status、Age、Num bathrooms 等信息。对影响用户的住房的状况的因素进行分析，得出影响因素的具体条件。

本文的主要流程如图 1。

2. 数据挖掘方法

数据挖掘(Data Mining)是通过分析每个数据，从大量数据中寻找其规律的技术，主要有数据准备、规律寻找和规律表示 3 个步骤。数据挖掘的任务有关联分析、聚类分析、分类分析、异常分析、特异群组分析和演变分析等。

3. 数据清洗

数据清洗是指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等[6-9]。数据清理是将数据库精简以除去重复记录，并使剩余部分转换成标准可接收格式的过程。数据清理标准模型是将数据输入到数据清理处理器，通过一系列步骤“清理”数据，然后以期望的格

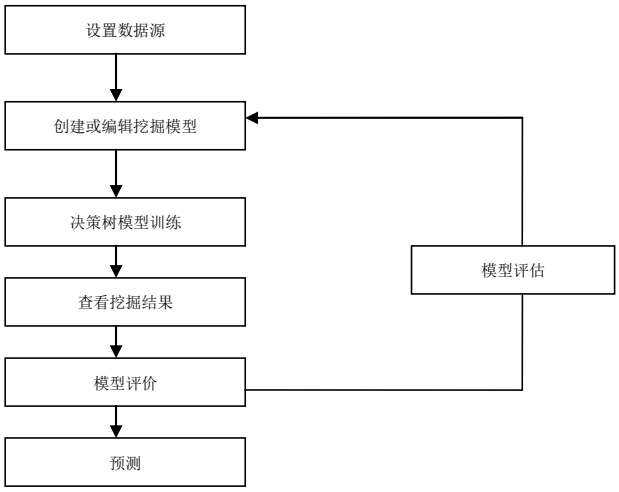

Figure 1. Flow chart
图 1. 流程图

式输出清理过的数据(如图 1 所示)。数据清理从数据的准确性、完整性、一致性、惟一性、适时性、有效性几个方面来处理数据的丢失值、越界值、不一致代码、重复数据等问题。

4. 决策树在预测中的应用

决策树是同时提供分类与预测的常用方法。决策树提供了一种展示类似在什么条件下会得到什么值这类规则的方法。比如，在贷款申请中，要对申请的风险大小做出判断，图是为了解决这个问题而建立的一棵决策树，从中我们可以看到决策树的基本组成部分：决策节点、分支和叶子[10]。每个分支要么是一个新的决策节点，要么是树的结尾，称为叶子。在沿着决策树从上到下遍历的过程中，在每个节点都会遇到一个问题，对每个节点上问题的不同回答导致不同的分支，最后会到达一个叶子节点。这个过程就是利用决策树进行分类的过程，利用几个变量(每个变量对应一个问题)来判断所属的类别(最后每个叶子会对应一个类别)。决策树技术是一种对海量数据集进行分类的非常有效方法。通过构造决策树模型，提取有价值的分类规则，帮助决策者做出准确的预测已经应用在很多领域。

4.1. 决策树算法具体分析

运行决策树算法之前，首先把输入的各项连续数据进行数据清洗，使其离散化[11]。决策树开始时，作为一个单个节点 N(根节点)包含所有的训练样本集，N 为图 2“全部”节点；决策树模型的预测项为经济损失比，其属性可以取 m 个不同的值，本文对经济损失比进行离散化，对应于 m 个不同类别为 Ci；设一个属性 A 取 v 个不同的值$\{a1, a2, \cdots, av\}$，若 A 取 Num bedrooms，则 A 取 4 个不同的值{Num bedrooms < 2, Num bedrooms = 2, Num bedrooms = 3, Num bedrooms > = 4}。利用属性 A 可以将 N 划分为 V 个子集$\{S1, S2, \cdots, Sv\}$，其中 Sj 包含了 S 集合中属性 A 取 aj 值的数据样本。若属性 A 被选为测试属性，设 sij 为子集 sj 中属于 Ci 类别的样本数。那么利用属性 A 划分当前样本集合所需要的信息(熵)可以计算如下：

$$E(A)=\sum_{j=1}^{v}\frac{s1j+s2j+\cdots+smj}{s}I(s1j,s2j,\cdots,smj)$$

其中 $\frac{s1j+s2j+\cdots+smj}{s}$ 项被当作为第 j 个子集的权值，它是由所有子集中属性 A 取 aj 值的样本数之和除以 S 集合中的样本总数。$E(A)$计算结果越小，就表示

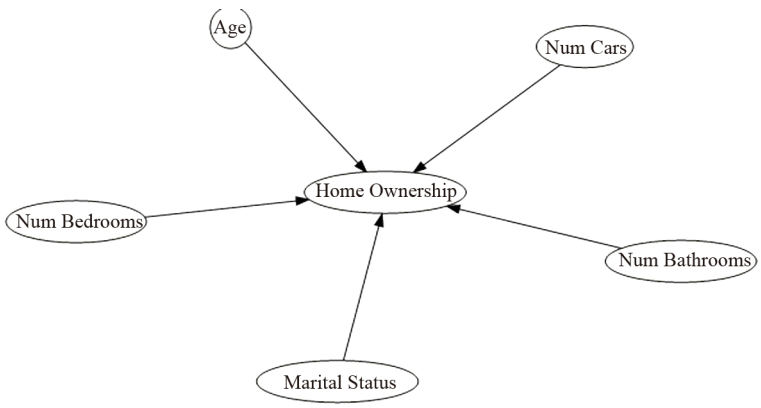

Figure 2. The dependency network
图 2. 依赖关系网络图

其子集划分结果越好。而对于一个给定子集 sj，它的信息为：

$$I(s1j, s2j, \cdots, smj) = -\sum_{i=1}^{m} pij \ \log(pij)$$

其中 $Pij = \frac{sij}{|Sj|}$，即为子集 sj 中任一个数据样本属于类别 Ci 的概率。

这样利用属性 A 对当前分支节点进行相应样本集合划分所获得的信息增益就是：

$$\text{Gain}(A) = I(s1, s2, \cdots, sm) - E(A)$$

也就是说，Gain (A)被认为是根据属性 A 取值进行样本集合划分所获得的(信息)熵的减少。本文中 Gain (Num bedrooms)、Gain (Num cars)、Gain (Marry Status)、Gain (Age)、Gain (Num bathrooms)等信息增长中，Gain (Num bedrooms)值最大，因此被作为测试属性用于产生当前分支节点，test_attribute = Num bedrooms。同时根据“Num bedrooms”取不同的值，把全部的输入分为 4 部分：Num bedrooms < 2, Num bedrooms = 2, Num bedrooms = 3, Num bedrooms ≥ 4，若设符合此条件的集合：Num bedrooms < 2 集合为 $s1$，返回值为 Generate_decision_tree ($s1$, Num bedrooms)。Num bedrooms = 2 集合为 $s2$，返回值为 Generate_decision_tree ($s2$, Num bedrooms)。Num bedrooms = 3 集合为 $s3$，返回值为 Generate_decision_tree ($s3$, Num bedrooms)。Num bedrooms > = 4 集合为 $s4$，返回值为 Generate_decision_tree ($s4$, Num bedrooms)。以此类推，继续递归调用决策树算法。

根据图 2 所示的训练样本集合，递归的使用上述各个处理过程；针对所获得的每个划分均又获得一个决策(子)树。一个属性一旦在某个节点出现，那么它就不能再出现在该节点之后所产生的子树节点中。

递归的按照上述步骤构造决策树，最终产生一个如图 3 所示的决策树。

4.2. 挖掘结果分析

如图 3 所示，每个矩形方框中不同颜色的直方图分别表示经济损失的不同损失程度的等级：蓝色直方图表示 owner 用户自己拥有住房，红色直方图表示 rent 用户自己租房子住。

八个叶子节点其中之一的叶子节点的具体挖掘图例如表 1 所示。

根据图的结果展示，当卧室数量 < 2 或是卧室数量 = 2 同时卫生间的数量 < 2.300 时，用户是租房子的概率比较大。反之，卧室数量 = 3 或是卧室数量 ≥ 4 时，用户是自己拥有房子的概率比较大，这点也

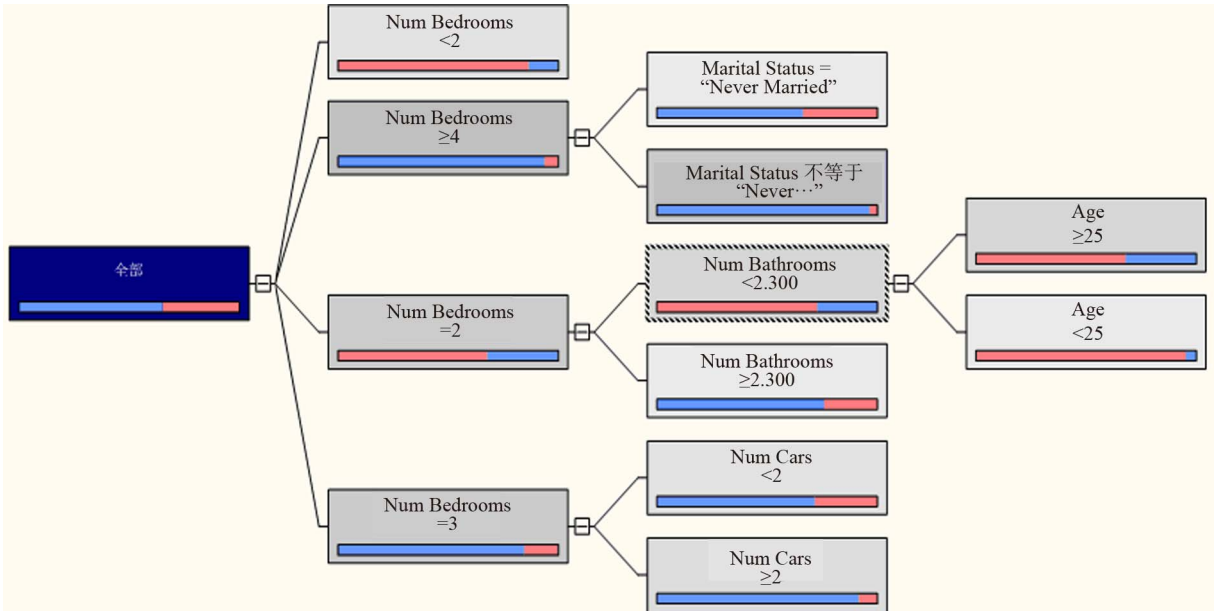

Figure 3. The result of decision tree
图 3. 决策树结果

Table 1. Num bedrooms < 2
表 1. 用户卧室的数量 < 2

值	事例	概率
Own	39	13.61%
Rent	250	86.39%
缺失	0	0.00%

比较符合常识，通常卧室的数量比较多，比较容易推测出来家庭成员比较多，所以比较倾向于自己拥有住房而不是租房。当卧室数量 ≥4 时，是否已婚对用户是否拥有自己的住房的影响比较大，为婚姻状况是未婚时，租房的概率会比拥有住房的概率有所增加。

依赖关系网络显示了模型中的输入属性和可预测属性之间的依赖关系。通过决策树算法分析依赖关系强度不同，可以得出对用户是否拥有自己的住房影响因素从弱到强依次为 Age (用户年龄)、Num cars (用户私家车的数量)、Num bathrooms (用户卫生间的数据)、Marital status (用户是否已婚)、Num bedrooms (用户卧室的数量)。这也与我们的基本认识相符合，一般情况下，年龄越大的人拥有房子的概率也会越大，拥有私家车数量比较多的人，说明家里人口比较多，经济情况比较好，所以越有可能有自己的住房等等。

5. 模型评估

本文研究中，挖掘结果的评价采用了微软的 Microsoft SQL Server 2008 模型评估评估模块，将挖掘结果导入到模型评估系统中，随机抽取了样本作为模型评估测试数据，对本研究结果做了准确性评估测试。其评估原理为从原数据里随即抽取 5%~33%测试，对其进行测试并且保持随机性。出现错误的预测与预测总数之间的比成为错误率。

如图 4 所示蓝色线代表理想模型的提升结果，表示在总体达到 100%时，总体的正确率也达到 100%，红色线代表决策树实际的提升结果(评估结果)。一般认为准确率砸 60%以上的模型都是具有一定的准确度，因此通过该模型得到的规则信息都能够帮助研究人员和采集者做出正确的决策。从图 5 中可以看出，决策树的分数为 0.89，预测的准确率达到 91.36%。其整体趋势还是跟理想模型的趋势比较相同，在预测评估住房情况方面，决策树模型的预测的效果比较好。

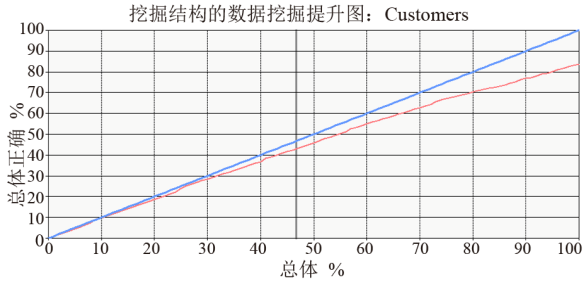

Figure 4. Model lift chart
图 4. 模型提升图

Figure 5. The model lift chart's mining legend
图 5. 模型提升图挖掘图例

6. 致谢

首先要感谢我的导师，给予我莫大的支持，在此深表感谢。然后，感谢平时帮助我的同学们，和你们的讨论中，我得到了成长，知识得到了巩固。最后，感谢本文引用文献、著作的作者，有了你们的无私奉献，才使得我们能站在巨人的肩膀上继续的探索。

参考文献 (References)

[1] J.-W. Han, M. Kamber, 著, 范明, 孟小峰, 译. 数据挖掘: 概念与技术[M]. 北京: 北京工业出版社, 2001: 3-4.
[2] 王丽珍, 周丽华, 陈红梅等. 数据仓库与数据挖掘原理及应用[M]. 北京: 科学出版社, 2005: 10-13.
[3] 王曰芬, 章成志, 张蓓蓓, 吴婷婷. 数据清洗研究综述[J]. 现代图书情报技术, 2007, 12: 50-56
[4] Pang-Ning Tan, Michael Steinbach, 著, 范明, 范宏建, 译. 数据挖掘导论[M]. 北京: 人民邮电出版, 2006: 35-40.
[5] 李云强. 数据挖掘及其在机车质量控制系统中的应用研究[D]. 西南交通大学, 2006.
[6] S. M. Weiss, C. A. Kulikowski. Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning, and expert systems. Burlington: M. Kaufmann Publishers, 1991.
[7] S. K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. Boston: Kluwer Academic Publishers, 1998.
[8] J.-W. Han, M. Kamber, 著, 范明, 孟小峰, 译. 数据挖掘: 概念与技术[M]. 北京: 北京工业出版社, 2001.
[9] 蔡国强, 贾利民, 吕晓艳, 刘春煌. 基于决策树的轨道交通安全评估方法及其应用[J]. 自然科学进展, 2007, 17(11): 1538-1543.
[10] M.-S. Chen J.-W. Han and P. S. Yu. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge

and Data Engineering, 1996, 8(6): 866-883.
[11] T. G. Dietterich. Overfitting and undercomputing in machine learning. ACM Computing Surveys—CSUR, 1995, 27(3): 326-327.

