改进的 C4.5 算法在绩效管理中的应用①

魏现波[1]，谢文阁[1]，王长斌[2]，张树奇[3]

[1](辽宁工业大学 电子与信息工程学院，锦州 121001)
[2](河北省邯郸市鸡泽县曹庄校区，邯郸 057350)
[3](长春工业大学 人文信息学院，长春 130012)

摘　要：提出了“得分变化率”和“部门权重”的定义，对 C4.5 算法进行改进；根据雪花模型构建了面向不同应用的基于绩效管理系统的数据仓库；用改进的 C4.5 算法挖掘数据仓库中的有用信息来确定绩效指标及指标权重，从而使考核结果更科学，结果分析更准确。

关键词：C4.5 算法；绩效管理；数据仓库；绩效指标；指标权重

Application of Improved C4.5 Algorithm in Performance Management

WEI Xian-Bo[1], XIE Wen-Ge[1], WANG Chang-Bin[2], ZHANG Shu-Qi[3]

[1](Electronic and Information Engineering College, Liaoning University of Technology, Jinzhou 121001, China)
[2](Jize Campus of Handan Hebei Province, Handan 057350, China)
[3](Human Communication College, Changchun University of Technology, Changchun 130012, China)

Abstract: This paper proposed the definition of “ScoreChangeRate” and “DepartmenWeight” and have improved the C4.5 algorithm. The data warehouse, based on the performance management system and applied to all kinds of aspect, is constructed according to the snowflake model. Using the improved of C4.5 algorithm abstract useful information in data warehouse to determine performance indicators and the weight can make the evaluation results more scientific and the results of analysis more accurate.

Key words: C4.5 algorithm; performance management; data warehouse; performance indicator; indicator weight

随着信息社会的不断发展，数据仓库、数据挖掘技术应用范围的不断广泛，并且越来越受到人们的关注。如何创新的应用数据仓库、数据挖掘，成为我们当前面临的主要问题。本文致力于数据挖掘算法的研究，对数据挖掘算法（C4.5）进行改进，提出“指标变化率”和“部门权重”的定义，并把改进的算法应用到基于数据仓库的绩效管理系统中，从而使指标制定和权重的确定更符合实际，使考核结果更科学，结果分析更准确。

1　绩效管理系统数据仓库设计

1.1 绩效管理系统数据仓库总体结构设计

根据需求分析阶段确定的主题以及最终要实现的应用需求，我们把绩效管理系统数据仓库的体系结构划分为三层（如图 1）：

(1) 数据源层[1]：从绩效管理关系型数据库中组织构建数据仓库所需关键数据。(2)数据处理与存储层 1：对数据源层组织的数据进行抽取、转换、清洗、装载，最终按照一定的规则构建成数据仓库及数据集市。(3)系统实现与业务应用层：实现数据仓库与绩效管理系统的衔接，根据数据仓库数据利用数据挖掘算法实现指标、任务及其权重的确定，形成考核计划，进行绩效评价。同时完成对结果数据的分析查询等具体应用。

1.2 绩效管理系统数据仓库的详细设计

(1) 主题确定　由于绩效管理是以被考评人为对象，以考核指标为主题，进行多维度评价。而对于我

① 基金项目:辽宁省教育厅研究项目(2008314)
收稿时间:2010-10-15;收到修改稿时间:2010-11-08

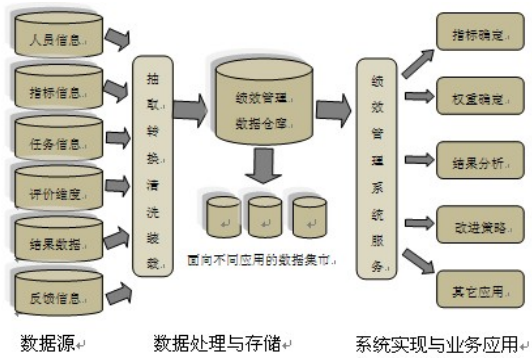

图 1 绩效管理系统数据仓库的体系结构

们挖掘分析处理最有价值的是评价数据，所以我们建立以被考评人、考核指标、考核任务、权重确定，考核结果、结果分析、改进策略等为主题的多级主题关联的主题及其之间的关系。

(2) 数据组织 根据绩效管理系统数据的特点本文采用雪花模型构建数据仓库。下面构建“指标及权重”（如图 2）主题的雪花模型组织数据。

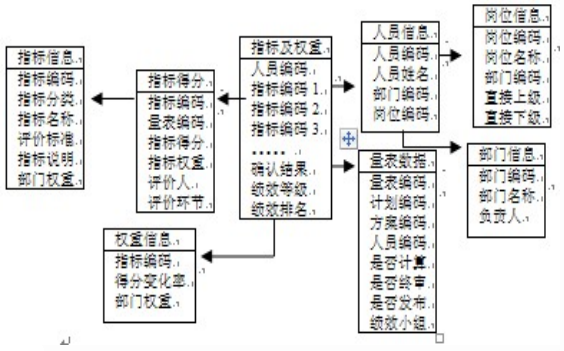

图 2 指标及权重雪花型结构图

2 C4.5算法研究及改进

2.1 基本概念定义[2]

设数据划分 S 为类标记的元组的训练集，假定类标号属性有 m 个不同的值，定义 m 个不同的类 C_i（i=1，i=2,…,m）.设 $C_{i,D}$ 是 D 中 C_i 类的元组集合，|D|和| $C_{i,D}$ |分别是 D 和 $C_{i,D}$ 中的元组个数。假设我们按指标 A 划分 D 中的元组，指标 A 具有 V 个不同的值{a_1，a_2，…,a_v}。用指标 A 将 D 划分为 v 个子集{D_1,D_2,…,D_v}。

(1) 信息增益 $Gain(A) = Info(D) - Info_A(D)$，其中，对 D 中元组分类所需要的期望信息 $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$ 也称为信息熵；为得到按指标 A 准确分类所需要的信息量 $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$。

(2) 增益率 $\text{GainRatio(A)} = \frac{\text{Gain(A)}}{SplitInfo(A)}$，

其中，分裂信息 $SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2(\frac{|D_j|}{|D|})$；

2.2 C4.5 算法介绍[3]

C4.5 算法是从 ID3 算法演变而来的，除了拥有 ID3 算法的功能外，C4.5 算法引入了新的方法和功能：（1）采用信息增益率度量，改变 ID3 算法采用信息增益度量偏袒具有较多值的属性的缺点；（2）合并具有连续属性的值，改变 ID3 算法只能处理离散属性的限制；（3）处理缺少属性的训练样本；C4.5 算法计算每个属性的信息增益率，并选取具有最高增益率的属性作为给定集合的测试属性。对被选择的测试属性创建一个节点，并以属性标记，对该属性的每个值创建一个分支，并据此划分样本。

2.3 C4.5 算法改进

虽然C4.5算法利用信息增益率度量很好的解决了偏袒具有较多值的属性，但是对于绩效管理中考核指标来说可能会出现孤立属性和属性值基本一致的情况，并且同一指标对于不同部门的关注程度并不完全相同，对此我们提出了“部门权重”的概念加以改进；另外，在绩效管理中一些指标由于相对稳定并且完成情况较好可能会具有较高的信息增益率，而对于一些波动性相对较大的指标，即完成情况不稳定信息增益率可能会比较低，但是这些波动性较大的指标可能正式绩效考核应该考核的指标，所以我们引入“得分变化率”来改进这一缺陷。改进定义如下：

得分变化率（ScoreChangeRate(A)）：相邻考核计划下同一指标得分变化率的平均值。

部门权重（DepartmenWeight(A)）：由部门负责人设定，不同指标相对于不同部门的重视程度。

由此，我们可以定义：

指标权重指数

$$Target\text{WeightIndex(A)} = \frac{GainRatio(A)}{\sum_{j=1}^{h} GainRatio(A_h)} * DepartmenWeigh(A) * ScoreChangeRate(A)$$

（假设一共有 h 个指标属性）。

指标的重要程度 MaterialityLevel(A)=GainRatio(A)* DepartmenWeigh(A)* ScoreChangeRate(A)

由于实际考核中指标权重采用百分比，所以我们定义：

指标权重

$$\text{TargetWeight(A)}=\frac{\text{TargetWeightIndex(A)}}{\sum_{j=1}^{h}\text{TargetWeightIndex}(A_j)}*100$$

根据 MaterialityLevel(A)来选择保留哪些指标，删除哪些指标。

3　应用分析

在第一章中我们已经根据构建了“指标及权重”这一主题的数据集市。我们从数据集市中抽取烟草公司销售部门的考核结果如图 3 所示。

序号	人员编码	销售额	销售费用比	完成比例	客户满意度	文档报告	绩效等级
1	01001	A	B	A	A	A	A
2	01002	B	B	B	B	B	B
3	01003	C	C	D	B	B	C
4	01004	C	A	D	B	D	C
5	01005	B	C	D	B	D	C
6	01006	A	B	B	B	C	B
7	01007	B	C	D	B	C	B
8	01008	B	A	A	A	A	A
9	01009	C	C	D	B	B	C
10	01010	B	C	B	A	C	B
11	01011	A	A	A	A	B	A
12	01012	C	A	D	B	C	B
13	01013	B	B	A	A	A	A
14	01014	C	A	D	A	B	B
15	01015	B	C	B	B	C	B
17	部门权重	0.90	0.85	0.95	0.50	0.40	
18	得分变化率	0.70	0.35	0.68	0.20	0.20	

图 3　销售部门考核结果表

(1) 计算信息熵 15 个销售人员的级别分别是 A、B、C 三种，有 4 个 A,7 个 B，4 个 C。则

$$Info(A,B,C)=\frac{4}{15}\log_2\frac{15}{4}+\frac{7}{15}\log_2\frac{15}{7}+\frac{4}{15}\log_2\frac{15}{4}=1.5324。$$

(2) 计算信息增益率

以销售额为例,信息熵：

$$Info_{\text{销售额}}(A,B,C)=\frac{3}{15}(\frac{2}{3}\log_2\frac{3}{2}+\frac{1}{3}\log_2 3)+\frac{7}{15}(\frac{2}{7}\log_2\frac{7}{2}+\frac{1}{7}\log_2 7)+\frac{5}{15}(\frac{2}{5}\log_2\frac{5}{2}+\frac{3}{5}\log_2\frac{5}{3})=1.152$$

信息增益：

$$Gain(\text{销售额})=Info(A,B,C)-Info_{\text{销售额}}(A,B,C)=0.3801-$$

分裂信息：

$$SplitInfo(\text{销售额})=-(\frac{3}{15}\log_2\frac{3}{15}+\frac{7}{15}\log_2\frac{7}{15}+\frac{5}{15}\log_2\frac{5}{15})=1.0438$$

信息增益率：

$$GainRatio(\text{销售额})=\frac{Gain(\text{销售额})}{SplitInfo(\text{销售额})}=0.3641$$

指标的重要程度：

$$\text{MaterialityLevel(销售额)}=\text{GainRatio(销售)})*\text{DepartmenWeigh(销售额)}*\text{ScoreChangeRate(销售)})=0.2294$$

=0.2294

指标权重指数：

$$\text{TargetWeightIndex(销售额)}=\frac{GainRatio(\text{销售额})}{\sum_{j=1}^{5}GainRatio(X_j)}*\text{DepartmenWeigh(销售额)}*\text{ScoreChangeRate(销售额)})=0.0851$$

同理得到其它指标各计算结果如下表 1 所示：

表 1　各指标相关参数计算结果

指标名称	销售额	销售费用比	完成比例	客户满意度	文档报告
GainRatio(A)	0.3641	0.3317	0.1353	0.8955	0.9675
MaterialityLevel(A)	0.2294	0.0987	0.1028	0.0896	0.0774
TargetWeightIndex(A)	0.0851	0.0366	0.0382	0.0332	0.0287

(3) 结果分析

由表 1 得到指标库保留的指标以及指标的权重如表 2：

表 2　改进后各指标权重及是否保留表

指标名称	销售额	销售费用比	完成比例	客户满意度	文档报告
指标权重	45%	20%	20%	15%	0
是否保留	是	是	是	是	否

如果不增加“部门权重”和“得分变化率”则得到的指标及权重如下表 3：

表 3　改进前各指标权重及是否保留表

指标名称	销售额	销售费用比	完成比例	客户满意度	文档报告
指标权重	15%	15%	0	35%	35%
是否保留	是	是	否	是	是

对表 2 和表 3 进行对比分析发现在实际情况中销售部门最关心的应该是销售额、销售费用比和完成比例而不是文档报告和客户满意度，由此可见我们加入“部门权重”和“得分变化率”作为制定指标库和指标权重是必要的符合实际情况的。

造成这种现象的主要原因在于：从图 3 的数据得知指标“客户满意度”的值基本上是“A”、“B”即它的得分波动性较小值比较集中，而 C4.5 虽然对 ID3 偏袒属性值较多的属性有所改进，但改进力度还不够；

(下转第 213 页)

```
zdlxsel=zdlx[sel.selectedIndex];//当前操作字段类型
switch (zdlxsel){
  case "CB"://动态形成类别的下拉列表
  var ss,htmlss;
ss= zdjb [document.frmCxset.zdm. selected Index].split(",");
    htmlss="<select name='in_lb' id='in_lb' size=10 'style='width:180px;height:190px; ' disabled= 'true'>";
    for(i=0;i<ss.length;i++)htmlss+="<option>" +ss[i]+"</option>";
    htmlss+="</select>";
    break;
    }
    ……
    cxtj+=sel.options[sel.selectedIndex].value;
  }
  ……
</ script >
```

最后把已组装的标准 SQL 语句提交服务器处理，完成后返回查询结果。由于篇幅所限，这里就不在给出服务器端处理代码。

3 结语

查询是管理信息系统中的一个十分重要的功能，其效率的高低，直接关系到整个系统的总体性能[3]。笔者通过以上设计，实现了一个多条件组合查询模块。该模块由一张查询配置表和前后台代码组成，后台代码只需要编写一个函数，前台页面则由 JavaScript 脚本语言来完成大部分功能，因而整个模块结构简单、使用方便、易于维护，无论.NET 平台还是 Java 平台都可以使用。实践证明此模块可灵活方便地实现对数据的任意查询，目前已在多个大型管理信息系统中采用，受到用户的一致好评。

参考文献

1 陈光柱,李志蜀.组合查询的组合算法.计算机工程与应用, 2003,33(1):197–198.
2 邵明.组合查询模块的设计与实现.计算机工程, 1999, 25(6):64–65.
3 叶春晓.数据库中多条件组合查询的方法及界面.重庆建筑高等专科学校学报, 1999,9(1):34–35.

(上接第 152 页)

而“完成比例”正好相反它的波动性较大，C4.5 计算机算出来的信息增益率就相对较小，而这种波动性较大的指标正是我们需要考核的；再有各个部门对不同的指标的重视程度也不相同。由此可见加入“得分变化率”和“部门权重”是非常必要的。

4 总结

在绩效管理中制定指标库和指标权重是绩效管理最重要的步骤，它直接关系到评价结果的有效性。文中在分析评价指标与评价对象各属性之间关系的基础上，对 C4.5 算法进行改进，在 C4.5 信息增益率的基础上提出了“部门权重”和“得分变化率”共同作为制定指标库和指标权重的阀值，从而使制定出来的指标及权重更科学、更符合实际要求。

参考文献

1 王丽珍,等.数据仓库与数据挖掘原理及应用.北京:科学出版社,2005.6–13.
2 毛国军等.数据挖掘原理与算法.第 2 版.北京:清华大学出版社,2007.20–60.
3 Han JW, Kamber M.数据挖掘概念与技术.第 2 版.北京:机械工业出版社,2007.
4 付亚和,许玉林.绩效考核与绩效管理.北京:电子工业出版社,2004.
5 谢辉.绩效管理中数据挖掘技术研究[硕士学位论文].武汉:华中科技大学,2006.