面向信用评级的有序决策树算法研究

裴生雷[1]，周伟[2]

（1.青海民族大学 物理与电子信息工程学院，西宁 810007；2.西宁市大数据服务管理局，西宁 810000）

摘要：决策树是数据挖掘领域的经典算法，应用领域非常广泛。在信用评价任务中，数据之间存在序关系，而传统的决策树算法无法解决这类问题。有序决策树的提出有效地解决了此类问题，能够从中发现新的知识，然而很多任务中属性与决策存在单调关系，并且样本之间无法比较，这影响了有序分类器的性能。因此，文章提出一种改进的有序决策树算法（Rank-DT）并应用于信用评价任务中，实验证明提出的算法改进了传统决策树算法的性能，获得了较好的效果。

关键词：序关系；决策树；信用评价；可比较的

中图分类号：TP181 **文献标志码**：A **文章编号**：1007-984X(2020)04-0009-05

在信用评价任务中，特征和决策是有序的，充分利用这些有序的特征信息能够获得潜在的价值。近几年，越来越多的研究者开始关注有序分类问题，这类任务广泛存在于现实世界。信用评价就是要利用有序特征中蕴含的信息来完成信用的等级评定。这类任务中需要考虑特征与决策之间的单调约束关系，也就是说信用评价等级依据特征值的大小递增或递减。银行客户信用评级对于衡量违约风险和违约概率是非常重要的。随着互联网技术的快速发展，信贷业务地开展更加灵活和高效，然而对于客户信用的评价体系的构建也提出了新的挑战。客户在银行信用系统中的信息量不断增加。基于先进的智能决策平台，辅助银行信贷业务更好地开展是非常必要的，进而降低风险和违约概率[1]。银行信用评价往往会从个人客户的年龄、职业、职位、个人收入、历史违约等方面考虑，而等级作为决策属性，取值分为 AAA 级，AA 级，A 级，BBB 级，BB 级，B 级，CCC 级。有序决策树模型在信用评价系统中发挥了重要的作用，能够有效的完成分类、排序。

这里，本文提出一种改进的有序决策树算法。该算法基于非负最小二乘法，学习属性的权重向量，在新的特征空间中完成数据的划分，使得做出的决策更为合理。同时，它可以获取更多的不确定信息，例如属性和决策的单调约束关系。这些都使得算法的效率大大提高，更精确地完成信用评价等级的衡量，辅助银行业务人员做出更好的决策。

1 决策树算法介绍

决策树在多标准决策领域是一种非常重要的工具，由于决策树能够将各种复杂的决策问题分解成简单的决策问题，为实际分类任务提供可靠的解决方案，决策者易于理解这些决策集合，因此受到广泛的关注和应用。在机器学习与数据挖掘领域，决策树也是一种非常重要的分类器。决策树应用递归分治策略，速度快、分类精度高。经典的决策树算法有 ID3，C4.5 以及 CART 树，这些决策树尽管泛化能力较好，也被成功应用于诸多领域，如图像识别、医疗诊断等。但是，这些算法并没有考虑属性值之间的序关系，对于属性与决策之间存在单调关系的决策问题无能为力，因此，提出一种改进的算法非常必要。提出的算法可以充分地挖掘潜在的序关系，并达到较好的分类效果。

2 有序分类问题

收稿日期：2020-03-20

基金项目：青海省应用基础研究项目（2019-ZJ-7017）；青海民族大学高层次项目（2020XJG13）；青海民族大学多源数据融合及应用科研创新团队

作者简介：裴生雷（1980-），男，山东潍坊人，副教授，博士，主要从事机器学习与数据挖掘研究，peishenglei@qhmu.edu.cn。

对于属性值之间存在序关系的单调分类任务，已经提出了一些算法。这些算法都是基于可比较的样本对假设前提下提出的，然而样本在某些属性上取值比另一个好，在其它属性上比另一个差，称之为不可比较的样本对。

因此，提出一种改进的有序决策树算法，处理信用评价系统中的不可比较样本对问题。给定有序分类样本集 $U=\{x_1,x_2,\cdots,x_n\}$，特征集 $A=\{a_1,a_2,\cdots,a_m\}$，$B\subseteq A$，Y 是一组类标记，并且有 y_i 是 x_i 的类标记。

2.1 样本之间的可比较性

在有序分类任务中，样本之间并不都是可以比较的，它可能在一些属性上取值比另一个好，而在其它属性上比另一个差，这种情况的存在影响着分类器的性能[2]。

定义 1 假设 $x,x'\in U$，如果 x 关于 A 占优于 x'，或者 x' 关于 A 占优于 x，那么 x 和 x' 是可比较的。如果 x 与 x' 不存在占优关系，那么就可以说 x 与 x' 是不可比较的。

现实世界中的很多任务都存在不可比较的样本，解决样本的不可比较问题能够获取更有价值的信息和知识。

2.2 单调一致性问题

在 U 上，如果样本对是可比较的，并且属性与决策之间的关系可以通过特定的单调函数来表示，这类问题中需要决策在 U 上达到单调一致性[3,4]。形式化定义如下：

定义 2 给定对象集 U，A 是属性集，$B\subseteq A$，假设 $\forall x_i,x_j\in U$，如果 $x_i\preceq_B x_j\Rightarrow Y(x_i)\leq Y(x_j)$ 或者 $x_i\succeq_B x_j\Rightarrow Y(x_i)\geq Y(x_j)$，那么我们说在 U 上的决策关于 B 是单调一致的，否则是非单调的。排序互信息受互信息的启发而提出，反映对象根据属性值提供的信息进行排序的一致性程度[5]。

假设给定有序数据集 $U=\{x_1,x_2,\cdots,x_n\}$，包含属性集 A，其中 $B\subseteq A$，$C\subseteq A$，数据集 U 关于 B 和 C 的排序互信息，形式化定义如下[5,6]。

向上的排序互信息：

$$RMI^{\succeq}(B,C)=-\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{\left|[x_i]_B^{\succeq}\right|\times\left|[x_i]_C^{\succeq}\right|}{|U|\times\left|[x_i]_B^{\succeq}\cap[x_i]_C^{\succeq}\right|}$$

向下的排序互信息：

$$RMI^{\preceq}(B,C)=-\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{\left|[x_i]_B^{\preceq}\right|\times\left|[x_i]_C^{\preceq}\right|}{|U|\times\left|[x_i]_B^{\preceq}\cap[x_i]_C^{\preceq}\right|}$$

其中，$|U|$ 表示全部样本数，$\left|[x_i]_B^{\succeq}\right|$ 和 $\left|[x_i]_C^{\succeq}\right|$ 分别表示 x_i 在属性集 B 和 C 上不比其它样本差的样本个数，$\left|[x_i]_B^{\succeq}\cap[x_i]_C^{\succeq}\right|$ 表示 x_i 在属性集 B 上不比其它样本差的样本与属性集 C 上不比其它样本差的样本交集的基数。同理可知，$\left|[x_i]_B^{\preceq}\right|$、$\left|[x_i]_C^{\preceq}\right|$ 以及 $\left|[x_i]_B^{\preceq}\cap[x_i]_C^{\preceq}\right|$ 的含义。

2.3 改进的有序决策树算法的提出

在有序分类问题中，单调分类作为一种特殊的分类问题而存在[7,8]。使用经典的决策树算法会产生非单调一致的决策规则，它们的存在直接关系到分类器的性能。在这里，不仅要考虑不可比较的样本对问题，还要保证产生单调一致的决策规则。因此，提出一种改进的有序决策树算法是非常必要的，以期在信用评价任务中的分类性能得到显著的提高。在该算法中，通过学习一系列的线性组合来完成不可比较到可比较样本对的转换，提出单调决策树算法。假定 $x,x'\in U,a_1,a_2\in A,a_1(x')<a_1(x)$ 并且 $a_2(x')>a_2(x)$，这里应用线性函数 $f(x)=w^Tx+b$ 来完成不可比较到可比较的转换，得到 $w^Tx'<w^Tx$ 或者 $w^Tx'>w^Tx$，显示两个属性的权重，使其可以进行比较。

利用函数 $f(x)$ 可以实现样本对的转换是显而易见的，获得有效的 w 是改进的有序决策树算法要解决的一个问题[9,10]。因此，为保证数据的单调一致性并产生有效的 w，给出目标函数的定义。对于给定的矩阵 $\boldsymbol{U}\in R^{n\times m}$ 以及 $\boldsymbol{y}$ 向量，目标函数为

$$\arg\min_{w}\left\|Uw+b-y\right\|_2^2,\ \ w\geq 0$$

根据计算出的 w，并形成 w 上的投影，进而在此基础上构建二叉树，产生单调一致的决策规则。

算法通过限制生成的权重非负，以保证数据的单调性。利用生成的权重向量 w，将数据集映射到新的数据空间，利用构建的单调决策树完成数据的分类。单调决策树的构建中比较关键的是找到一个合适的分裂点，将数据点划分为左右两部分。对象 x_i 在向量 w 上的投影用于形成可能的分裂点。

$$p_i=\frac{w^Tx_i}{\left\|w\right\|},i=1,2,\cdots,n$$

每一个分裂点通过 p_{test} 完成数据集的划分。

$$p_{\text{test}}=\frac{w^Tx_i}{\left\|w\right\|},i=1,2,\cdots,n-1$$

每一个划分依据分裂准则形成最好的分裂点，如果 $p_i\leq p_{\text{test}}^{*}$，那么左分区 $part_{\text{left}}=\left\{x\in U_i,\frac{w^Tx}{\left\|w\right\|}\leq p_{\text{test}}\right\}$；否则，右分区 $part_{\text{right}}=\left\{x\in U_i,\frac{w^Tx}{\left\|w\right\|}>p_{\text{test}}\right\}$，这个过程直到停止规则满足时才会终止。

算法扫描所有的数据点并计算当前节点向下的排序互信息，从中找出最大的一个，确定为最好最合理的分裂点，不断递归，最终产生决策规则。

改进的有序决策树算法（Rank-DT）为

输入：训练样本集合，样本用（A,D）来表述

停止参数：ε=0.01，L=1；

输出：有序决策树。

1 生成决策树根节点；

2 如果剩余的样本数小于等于 L 或者所有的样本属于同一类，则标记为叶节点，返回；
　否则，执行步骤 3；

3 计算不同属性的权重 w；
　　计算样本在向量 w 上的投影；
　　计算排序互信息的值，并得到每次的最大值作为最好分裂点；

4 选择所有分裂点中的最大值；

5 如果最大值小于 ε
　　标记叶节点，返回；
　否则，继续执行步骤 6；

6 最好的分裂点划分父节点为左右子节点；

7 依据左右子节点，递归的构造有序决策树。

3　算法在信用评价中的应用

信用评价模型在商业银行中具有举足轻重的地位，通过信用评价或评级来检验风险规律和预测风险。提出的算法主要用于信用评级，通过对银行客户的合理排序，可以进一步计量风险，估计客户的违约概率。

3.1　信用评价体系的关键

信用评价体系中的关键是用尽可能少的变量准确预测等级及风险，如果能够找到一个变量来预测是最为理想的。然而，实际应用中不可能存在这个变量，选择合适的变量做出预测就是一个值得考虑的问题。因此，在构建计算模型时需要充分地利用数据特性找出用户风险特征，挖掘潜在用户，同时在审核和授信过程中加以控制，尽最大可能降低信用风险，防止因信用问题导致的资金损失。

例如，银行信用卡系统中用户逾期数据可以作为信用评价重要依据，根据客户逾期记录的关键属性，

如性别、信用卡使用率、信用卡额度、住房贷款、历史逾期行为、开户行为等，对客户的信用状况进行评级。通过构建合适的模型和算法理解客户是否容易发生逾期行为，以及哪些因素会影响客户的信用等级或者是逾期的严重性程度，为新客户办卡提供重要参考信息。

3.2 信用评价任务中的实验分析

利用提出的算法分别在UCI中的3个信用评价相关数据集bankruptcy risk，German Credit以及Credit Approval上进行相关实验。通过性能评价指标判断在考虑单调约束以及样本对不可比较的问题的情况下，算法的性能是否优于其它的传统算法和有序分类算法[11]。

数据集bankruptcy risk涵盖了39家公司的风险评估数据，通过12个属性来描述；German Credit抽取了1 000个样本，其中700个信用评价较好，300个信用评价较差，通过20个属性来描述。Credit Approval数据来自于信用卡系统，抽取了690个样本以及15个属性。

通过训练的决策树完成信用评级，并分析算法是否对信用等级有良好的区分能力。实验中选取了4个算法：OSDL、OLM、J48和CART，其中OSDL和OLM是有序分类算法，J48和CART是一般分类算法[9,12,13]。提出的算法为有序决策树，简称Rank-DT。在实验中，应用十折交叉验证技术验证信用等级的平均结果，具体通过两个个评价指标完成性能比较，分别是分类正确率（PCC）和平均绝对误差（MAE）。这里对3个评价指标加以定义和说明，已明确其有效性。假定数据集的样本总数为 N，y_n 为样本的实际类别，y_n' 为预测的类别，那么分类正确率定义为

$$\mathrm{PCC}=\frac{1}{N}\sum_{n=1}^{N}\left(\hat{y}_n\oplus y_n\right)$$

其中，当 $\hat{y}_n=y_n$ 时，$\hat{y}_n\oplus y_n=1$，否则 $\hat{y}_n\oplus y_n=0$。平均绝对误差的定义为

$$\mathrm{MAE}=\frac{1}{N}\sum_{n=1}^{N}\left|\hat{y}_n-y_n\right|$$

对于bankruptcy risk任务，实验结果如表1所示。可以清楚的看到，Rank MMT算法在性能指标上优于其它算法。这得益于该算法能够更好地处理不可比较样本对的问题，例如A公司的投资回报率比B公司的投资回报率高，那么A公司的风险等级可能比B公司低。当然也可能会存在这样的情况，尽管A公司的投资回报率高于B公司，但是两个公司所在城市不同，导致两个A公司的风险更高。

表1 bankruptcy risk任务上不同算法的性能评价指标

	Rank-DT	OSDL	OLM	J48	CART
PCC	0.90	0.71	0.60	0.80	0.87
MAE	0.10	0.29	0.53	0.20	0.13

在表2中显示了Rank-DT算法在任务German Credit上的实验结果。根据统计检验确定不同算法的平均性能是否存在显著性差异，应用 t 检验两两比较了所有算法的平均性能，进而明确了提出算法的在不同性能评价指标上是否存在优势[14]。结果显示，Rank-DT算法性能优势不明显，非单调噪声的影响了算法的性能。

表2 German Credit任务上不同算法的性能评价指标

	Rank-DT	OSDL	OLM	J48	CART
PCC	0.70	0.51	0.69	0.70	0.70
MAE	0.30	0.49	0.31	0.30	0.30

针对信用卡系统中客户信息，可以直接来确定客户信用等级，从而为规避风险提供决策依据。Credit Approval是信用卡系统中的数据，Rank-DT算法在该任务上的实验结果显示在表3中，可以清楚地看到性能评价指标的差异，根据单尾 t 检验来确定算法之间在不同的性能指标上是否存在显著性差异[14]。最终，确定提出的算法在评价指标上的性能是显著的。

表3 Credit Approval任务上不同算法的性能评价指标

	Rank-DT	OSDL	OLM	J48	CART
PCC	0.88	0.80	0.78	0.86	0.85
MAE	0.12	0.20	0.22	0.14	0.14

通过以上的实验分析，可以推断Rank-DT算法在信用评价和风险评估任务中的应用效果，可以为不同的组织机构提供决策支持。

4 结束语

基于信用评价中存在的有序分类问题提出了一种改进的有序决策树。由于在这类任务中属性和决策是有序的，并且两者之间存在单调关系。属性与决策的单调关系可以使我们获取更多的知识以提供更为可靠的决策支持。文章提出的有序决策树算法 Rank-DT，在信用评价任务中效果显著，能够较好地应用于实际环境，为有序分类任务的实施提供一个合理的算法。

参考文献：

[1] Xia Y, Liu C, Li Y Y, et al. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring[J]. Expert Systems with Applications, 2017, 78: 25-241

[2] Kadzinski M, Greco S, Slowinski R. Robust ordinal regression for Dominance-based Rough Set Approach to multiple criteria sorting[J]. Information Sciences, 2014, 283: 211-228

[3] Augeri M G, Cozzo P, Greco S. Dominance-based rough set approach: An application case study for setting speed limits for vehicles in speed controlled zones[J]. Knowledge-Based Systems, 2015, 89: 288-300

[4] Marsala C, Petturiti D. Rank discrimination measures for enforcing monotonicity in decision tree induction[J]. Information Sciences An International Journal, 2015, 291(C): 143-171

[5] Hu Q H, Guo M Z, Yu D R, et al. Information entropy for ordinal classification[J]. Science China Information Sciences, 2010, 53(06): 1188-1200

[6] Hu Q H, Che X, Zhang L, et al. Rank entropy based decision trees for monotonic classification [J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24(11): 2052-2064

[7] 陈建凯，王熙照，高相辉. 改进的基于排序熵的有序决策树算法[J]. 模式识别与人工智能，2014, 27(02): 134-140

[8] 潘巍巍，宋彦萍，于达仁. 齿轮裂纹程度识别的有序分类算法[J]. 哈尔滨工业大学学报，2016, 48(07): 156-162

[9] Olson D L, Delen D. Advanced data mining techniques [D]. Springer Berlin Heidelberg, 2008

[10] 唐耀先，余青松. 消除属性间依赖的 C4.5 决策树改进算法[J]. 计算机应用与软件，2018, 35(03): 262-265, 315

[11] UCI. UCI machine learning repository, http://archive.ics.uci.edu/ml/

[12] Lievens S, Baets B D, Cao-Van K. A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting[J]. Annals of Operations Research, 2008, 163(01): 115-142

[13] WEKA. Machine learning group at university of waikato, http://www.cs.waikato.ac.nz/ml/weka/

[14] Demsar J. Statistical comparisons of classifiers over multiple data sets[J]. Journal of Machine Learning Research, 2006, 7(01): 1-30

Research on ordinal decision trees algorithm for credit rating

PEI Sheng-lei[1]，ZHOU Wei[2]

(1.College of Physics & Electronic Information Engineering, Qinghai University for Nationality, Xining 810007, China;
2. Xining big data Service Management Bureau, Xining 810000, China)

Abstract: Decision trees are a kind of classic algorithm in the data-mining field, and the application is very extensive. In the credit evaluation tasks, it exists an ordinal relationship in the data, and the traditional decision tree algorithm cannot solve such problems. The ordinal decision trees are effective to solve these problems, and it can discover new knowledge. However, there are monotonic relationships between attributes and decisions on many tasks. Moreover, it be compared among samples, which affect the performance of ordinal classifiers. Therefore, this paper proposes an improved decision tree algorithm (Rank-DT) and applies it to the credit evaluation tasks. Experiments results show that the proposed algorithm improves the performance of the traditional decision tree algorithm and obtains good effect.

Key words: ordinal relationship; decision trees; credit evaluation; comparable