研究简报

doi: 10.13413/j.cnki.jdxblxb.2017.05.42

数据挖掘中改进的C4.5决策树分类算法

王 文 霞

(运城学院 计算机科学与技术系, 山西 运城 044000)

摘要: 针对传统C4.5决策树分类算法需要进行多次扫描, 导致运行效率低的缺陷, 提出一种新的改进C4.5决策树分类算法. 通过优化信息增益推导算法中相关的对数运算, 以减少决策树分类算法的运行时间; 将传统算法中连续属性的简单分裂属性改进为最优划分点分裂处理, 以提高算法效率. 实验结果表明, 改进的C4.5决策树分类算法相比传统的C4.5决策树分类算法极大提高了执行效率, 减小了需求空间.

关键词: 数据挖掘; C4.5决策树; 分类算法; 判别能力度量; 连续属性

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5489(2017)05-1274-04

Improved C4.5 Decision Tree Classification Algorithm in Data Mining

WANG Wenxia

(*Department of Computer Science and Technology, Yuncheng University, Yuncheng* 044000, *Shanxi Province, China*)

Abstract: Aiming at the problem that the algorithm for traditional C4.5 decision tree classification algorithm needed to be scanned several times, resulting in defects of running low efficiency, the author proposed a new improved C4.5 decision tree classification algorithm by optimizing the logarithmic operation related information gain derivation algorithm in order to reduce the running time of the decision tree classification algorithm. And the simple split attribute of the continuous attributes in the traditional algorithm was improved to the optimal partition point splitting processing in order to improve the efficiency of the algorithm. Experimental results show that compared with the traditional C4.5 decision tree classification algorithm, the improved C4.5 decision tree classification algorithm greatly improves the execution efficiency and reduces the demand space.

Key words: data mining; C4.5 decision tree; classification algorithm; discriminative ability measure; continuous attribute

随着数据库技术的普及推广, 数据挖掘越来越受到人们的广泛关注[1]. 数据挖掘就是将收集的海量数据先通过特定的分类算法进行分类, 然后对这些不同类别的数据进行分析对比, 并构建特定的数据模型综合分析这些不同类别的数据, 最后得出有价值的数据信息或通过数学建模对以后的数据进行合理的预测[2].

文献[3]利用决策树分类算法研发了相应的应用系统——CLS. 一种优秀决策树分类算法的特点是产生最少量的、深度最小的叶子, 且应用这种决策树分类算法分析海量数据的精度尽量高; 同时决策树不能太复杂, 太复杂会加重运算负担. Quinlan[4]提出了ID3决策树分类算法, 该算法将海量数据进行分类预测, 取得了较好的效果, 已成为一种经典的决策树分类算法, 但不能有效处理具有连续属

收稿日期: 2016-10-25.

作者简介: 王文霞(1979—), 女, 汉族, 硕士, 讲师, 从事算法分析的研究, E-mail: wangwx@126.com.

基金项目: 国家自然科学基金(批准号: 11241005)和山西省运城学院131人才专项基金(批准号: JG201634).

性的数据；Prasad 等[5]提出了分类与回归树算法，该算法对数据价值进行判断，能对无利用价值的数据进行忽略处理，但判断精度较低；Quinlan[6]提出了 C4.5 决策树分类算法，该算法有效改进了 ID3 决策树算法的缺陷，并能对某些数据的缺失进行有效处理，但算法建树过程复杂；为了改进 C4.5 决策树分类算法建树复杂的缺点，文献[7]提出了 SLIQ 分类算法，该算法建树过程简单，预测精度高，但算法受主存容量影响明显；文献[8]则提出一种改进的 SLIQ 分类算法——SPRINT 算法，该算法有效解决了 SLIQ 分类算法受主存容量影响的缺陷；PUBLIC 算法[9]有效解决了上述算法中决策树必须先建好再修剪的缺陷，极大提高了算法运行效率.

本文在 C4.5 决策树分类算法的基础上，提出一种改进的决策树分类算法. 该算法改进了 C4.5 决策树分类算法的判断能力度量公式，优化了对数运算，并在处理具有连续属性的海量数据时改进了 C4.5 决策树分类算法中最优划分点的选择，极大提高了分类算法的运行效率.

1　C4.5 决策树分类算法

C4.5 决策树分类算法是在 ID3 算法的基础上进行的改进，其核心为“信息熵”，通过信息增益比率量化属性判别能力. 信息增益比率表达式为

$$\mathrm{GainRatio}(A)=\frac{\mathrm{InfoGain}(A)}{I(|S_1|/|S|,|S_2|/|S|,\cdots,|S_n|/|S|)}. \tag{1}$$

虽然信息增益比率能有效解决 ID3 算法的缺陷，但比率数值有可能出现分母无限接近于 0 甚至为 0 的情况.

C4.5 决策树分类算法在处理数据时会针对不同属性的数据采用不同的处理方式. 在处理连续属性数据时，C4.5 决策树分类算法将连续属性数据先进行离散处理，并对数据段进行排序，然后分批处理不同片段的数据，效率较低. 虽然 C4.5 决策树分类算法较 ID3 算法提高了算法效率，但仍会生成大量的多叉树，而在建树过程中又必须对训练集进行扫描与排列处理，所以效率仍然未达到理想状态，且 C4.5 决策树分类算法并不能有效处理训练集体积超过主存容量的情况.

2　改进的 C4.5 决策树分类算法

2.1　*属性判别能力计算方法的优化*　本文针对 C4.5 决策树分类算法运行效率进行改进，对计算过程中的一些对数运算进行优化精简处理.

假设数据集 D 中只含有两种属性：大小为 y 的正例集 P_D；大小为 n 的反例集 N_D. 一棵决策树能对数据集做出合理判断的前提是必须掌握一定的信息量，掌握信息量计算公式为

$$\mathrm{Info}(D)=-\frac{y}{n+y}\log_2\frac{y}{n+y}-\frac{n}{n+y}\log_2\frac{n}{n+y}. \tag{2}$$

假设一棵决策树的根节点为属性 X，$X=(X_1,X_2,\cdots,X_s)$，通过 X 将数据集 D 分为 s 个数据子集，$D=\{D_1,D_2,\cdots,D_s\}$，每个数据子集中的数据属性都为相同标号的属性 X，并均含有正例集 P_D 和反例集 N_D，例如：D_1 数据的属性为 X_1，分别为数目为 y_1 的正例集 P_D 和数目为 n_1 的反例集 N_D，以此类推，则数据子集期望信息计算公式为

$$I(D_i)=-\frac{y_i}{n_i+y_i}\log_2\frac{y_i}{n_i+y_i}-\frac{n_i}{n_i+y_i}\log_2\frac{n_i}{n_i+y_i}. \tag{3}$$

由式(2)和式(3)可知，数据集 D 根节点为属性 X 时的信息熵量化公式为

$$\mathrm{Info}(D_i)=\sum_{i=1}^{v}\frac{y_i+n_i}{y+n}I(D_i)=\frac{1}{(n+y)\ln 2}\sum_{i=1}^{v}\left(-y_i\ln\frac{y_i}{n_i+y_i}-n_i\ln\frac{n_i}{n_i+y_i}\right), \tag{4}$$

式(4)中$\frac{1}{(n+y)\ln 2}$为一定值，为了简化计算过程可将其省略，对最终实验结果影响较小，则

$$\mathrm{Info}(D_i)=\sum_{i=1}^{v}\left(-y_i\ln\frac{y_i}{n_i+y_i}-n_i\ln\frac{n_i}{n_i+y_i}\right). \tag{5}$$

又因 $\ln(1+x)\approx x$，所以有

$$\ln\frac{y_i}{y_i+n_i}=\ln\left(1-\frac{n_i}{y_i+n_i}\right)\approx\frac{n_i}{y_i+n_i}, \tag{6}$$

$$\ln\frac{n_i}{y_i+n_i}=\ln\left(1-\frac{y_i}{y_i+n_i}\right)\approx\frac{y_i}{y_i+n_i}. \tag{7}$$

从而决策树能对数据集做出合理判断的前提为

$$\mathrm{Info}(D)=-\frac{y}{n+y}\log_2\frac{y}{n+y}-\frac{n}{n+y}\log_2\frac{n}{n+y}\approx\frac{2}{\ln 2}\frac{yn}{(n+y)^2}; \tag{8}$$

数据集 D 根节点为属性 X 时的信息熵量化公式为

$$\mathrm{Info}(D_i)=\sum_{i=1}^{v}\left(-y_i\ln\frac{y_i}{y_i+n_i}-n_i\ln\frac{n_i}{y_i+n_i}\right)\approx\sum_{i=1}^{v}\frac{2y_in_i}{y_i+n_i}; \tag{9}$$

分裂信息度量公式为

$$H(D_i)=-y_i\log_2\frac{y_i}{y_i+n_i}-n_i\log_2\frac{n_i}{y_i+n_i}\approx\frac{2}{\ln 2}\frac{y_in_i}{(n_i+y_i)^2}; \tag{10}$$

信息增益率计算公式为

$$\mathrm{GainRatio}(D_i)=\frac{\mathrm{Info}(D)-\mathrm{Info}(D_i)}{H(D_i)}. \tag{11}$$

改进的 C4.5 决策树分类算法采用新的经过改进的信息熵 $\sum_{i=1}^{v}\frac{2y_in_i}{y_i+n_i}$，并利用$\frac{2}{\ln 2}\frac{y_in_i}{(n_i+y_i)^2}$计算出分裂信息度量，优化了信息增益率计算公式，极大提高了算法运行效率.

2.2 连续属性数据处理优化 C4.5 决策树分类算法采用离散的方法对连续属性数据进行处理，将离散后的数据段中间数据点作为数据划分点计算信息增益率大小. 当某数据段中的数据均为连续属性时，则会导致 C4.5 决策树分类算法每次离散处理连续属性数据都需计算最优判别能力度量，其中包括计算一些没必要的划分点在内的全部划分点的信息增益率，导致算法运行效率较低.

研究表明，相同属性计算出的信息增益率也相同. 由边界点定义可知，连续属性最优划分点必出现在两个属性不同的数值之间，且两个数据彼此相邻[10]. 基于此，本文在连续属性数据处理中去除一些没必要的划分点的信息增益率计算.

设训练集 S 中的数据共有 c 种属性，通过排序方式将分裂后的连续属性 A 进行处理，则连续属性 A 具有$(n+l)$种不同的数据段，即具有 n 个不用的数据划分点，在实际应用中 $n\gg c$. 改进的 C4.5 决策树分类算法只计算最优划分点的信息增益率，排序好的连续属性 A 各片段中具有相同属性的片段相邻，它们的信息增益率也相同. 因此，改进的 C4.5 决策树分类算法只需计算$(c-l)$个划分点，从而加快了算法的运行效率.

3 实验结果与分析

本文实验数据为 UCI 机器学习数据库中 4 个子数据库的数据. 为了使结果尽可能准确，共进行 10 次传统 C4.5 决策树分类算法与改进 C4.5 决策树分类算法间的算法运行效率及运行时间对比交叉实验.

首先选取 Iris 数据集，该数据集 150 个样本的 4 个分裂属性都是连续属性，取 3 个类别属性. 用该数据集进行算法对比仿真实验，结果如图 1 所示. 由图 1 可见，当数据集样本数量较少，且属性全为连续属性时，改进的 C4.5 决策树分类算法在运行时间上较传统 C4.5 决策树分类算法少许多，并且精度未受影响，算法得到了大幅度优化.

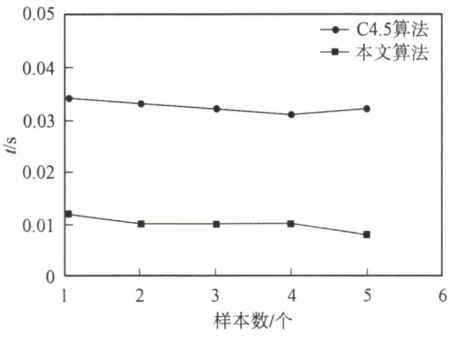

图 1 Iris 数据集算法改进前后的运行时间对比

Fig. 1 Running time comparisons of Iris data set before and after algorithm improvement

其次选取 Abalone 数据集，该数据集含有 3 133 个样本，属于大容量数据集，共 7 个连续属性与1 个分裂属性，取 3 个类别属性. 用该数据集进行算法对比仿真实验，结果如图 2 所示. 由图 2 可见，当数据集样本数量巨大时，算法运行时间较

使用 Iris 数据集进行算法实验时要长，并且改进 C4.5 决策树分类算法的运行时间相较传统 C4.5 决策树分类算法要少很多，算法得到了大幅度优化. 因此，数据样本容量越大，算法运行时间的差别越大.

最后选取 Vote 数据集进行算法仿真实验，结果如图 3 所示. Vote 数据集含有 300 个样本，有 16 个分裂属性，全部为离散属性. 由图 3 可见，具有离散属性的 Vote 数据集自身也较小，算法运行时间较其他数据集少. 改进 C4.5 决策树分类算法的运行时间较传统 C4.5 决策树分类算法相差不大. 因此，数据样本容量越小，算法运行时间的差别越小.

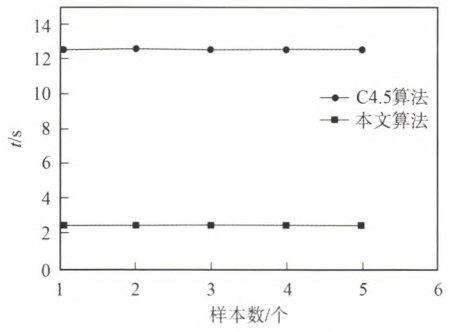

图 2　Abalone 数据集算法改进前后的运行时间对比

Fig. 2　Running time comparisons of Abalone data set before and after algorithm improvement

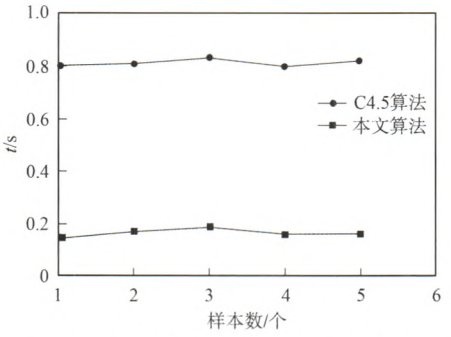

图 3　Vote 数据集算法改进前后的运行时间对比

Fig. 3　Running time comparisons of Vote data set before and after algorithm improvement

综上所述，本文针对传统 C4.5 决策树分类算法需要进行多次扫描，导致运行效率较低的缺陷，提出了一种新的改进 C4.5 决策树分类算法. 通过优化信息增益推导算法中相关的对数运算，极大减小了决策树分类算法的运行时间，并将传统 C4.5 决策树分类算法中连续属性的简单分裂属性改进为最优划分点分裂处理，提高了算法效率.

参　考　文　献

[1] 魏晓辉，李聪，李洪亮，等. 支持大规模流数据处理的在线 MapReduce 数据传输机制 [J]. 吉林大学学报(理学版)，2015，53(2)：273-279. (WEI Xiaohui，LI Cong，LI Hongliang，et al. Online MapReduce Data Transmission Mechanism Supporting Large-Scale Stream Data Processing [J]. Journal of Jilin University (Science Edition)，2015，53(2)：273-279.)

[2] 刘杰，刘大有，金弟. 一种基于模糊 C 均值的新分类算法 [J]. 吉林大学学报(理学版)，2009，47(4)：795-799. (LIU Jie，LIU Dayou，JIN Di. A New Classification Algorithm Based on Fuzzy C-Means [J]. Journal of Jilin University (Science Edition)，2009，47(4)：795-799.)

[3] Kira K，Rendell L. The Feature Selection Problem：Traditional Methods and a New Algorithm [C]//Proceedings of the 9th National Conference on Artificial Intelligence. Palo Alto：AAAI Press，2008：129-134.

[4] Quinlan J R. Simplifying Decision Trees [J]. International Journal of Human-Computer Studies，1986，27(3)：221-234.

[5] Prasad A M，Iverson L R，Liaw A. Newer Classification and Regression Tree Techniques：Bagging and Random Forests for Ecological Prediction [J]. Ecosystems，2006，9(2)：181-199.

[6] Quinlan J R. C4.5：Programs for Machine Learning [M]. San Francisco：Morgan Kaufmann Publishers，1993.

[7] Mehta M，Agrawal R，Rissanen J. SLIQ：A Fast Scalable Classifier for Data Mining [C]//International Conference on Extending Database Technology. Berlin：Springer，1996：18-32.

[8] Ramdane C，Meshoul S，Batouche M，et al. A Quantum Evolutionary Algorithm for Data Clustering [J]. International Journal of Data Mining Modelling & Management，2017，2(4)：369-387.

[9] Haraty R A，Dimishkieh M，Masud M. An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data [J]. International Journal of Distributed Sensor Networks，2015，2015：1-11.

[10] Liu K E，Lo C L，Hu Y H. Improvement of Adequate Use of Warfarin for the Elderly Using Decision Tree-Based Approaches [J]. Methods of Information in Medicine，2013，53(1)：47-53.

(责任编辑：韩　啸)