第 18 卷　第 16 期　2018 年 6 月
1671—1815(2018)016-0237-04

科　学　技　术　与　工　程
Science Technology and Engineering

Vol. 18　No. 16　Jun. 2018
ⓒ　2018　Sci. Tech. Engrg.

引用格式:高洪涛, 陆　伟, 杨余旺. 基于统计学法则的连续属性值划分方法[J]. 科学技术与工程, 2018, 18(16): 237—240
Gao Hongtao, Lu Wei, Yang Yuwang. Partition approach for continuous attributes based on statistical criterion[J]. Science Technology and Engineering, 2018, 18(16): 237—240

基于统计学法则的连续属性值划分方法

高洪涛[1]　陆　伟[2]　杨余旺[2]

(中国刑事警察学院网络犯罪侦查系[1], 沈阳 110035;南京理工大学计算机科学与工程学院[2], 南京 210094)

摘　要　目前决策树中很多分类算法例如 ID3/C4.5/C5.0 等都依赖于离散的属性值,并且希望将它们的值域划分到一个有限区间。利用统计学法则,提出一种新的连续属性值的划分方法;该方法通过统计学法则来发现精准的合并区间。另外在此基础上,为提高决策树算法分类学习性能,提出一种启发式的划分算法来获得理想的划分结果.在 UCI 真实数据集上进行仿真实验.结果表明获得了一个比较高的分类学习精度、与常见的划分算法比较起来有很好的分类学习能力。

关键词　连续属性值　　学习精度　　统计学法则　　分类算法

中图法分类号　TP393.03;　　**文献标志码**　A

目前决策树中很多分类算法例如 ID3/C4.5/C5.0 等都依赖于离散的属性值,并且希望将它们的值域划分到一个有限区间,这样就可以产生一个相对小数目的可区分的属性值,所以连续属性值的划分方法至关重要[1]。一个良好的连续属性值划分方法不仅要能够产生一个关于连续属性值的简化数据来帮助专家或者用户来分析数据,而且需要使数据的归纳学习过程更快与更准确,同时减少信息的丢失[2]。

文献[3—5]中描述了国内外当前的连续属性值的划分方法的总体描述,都已经讨论得比较全面了。其主要包括自底向上[6,7](bottom-up)和自顶向下[8,9](top-down)的属性值划分方法;有监督(supervised)和无监督(unsupervised)的属性值的划分方法[10]。

有监督方法考虑类标签信息,它相对于其他的方法来说更加成熟,例如基于熵(entropy-based)的属性值划分方法,基于类-属性相互依赖的属性值划分方法和基于卡方统计独立性的属性值划分方法[7]。

基于熵的属性值划分方法它选择使整体熵最小的断点(cut point)作为结果断点,并且通过最小描述长度原则(minimum description length principle, MDLP)来确定离散区间数[11];近年来,许多学者还研究了新型的属性值划分方法,例如半监督学习属性值划分方法[12],提取属性值划分方法[13],还有其他的相关的属性值划分相关的技术[14],总体来讲这些连续属性值的划分方法各有优点与缺点。

针对上述分析,本文着重研究自底向上 Bottom-up 的这类的属性值划分方法,提出一个新的划分方法,它使用到了统计学法则。该方法通过建立统计学法则来发现精准的合并区间,另外在此基础上,为改善决策树算法的分类性能,提出一种启发式的划分算法来获得理想的划分结果。最后在 BenchMark 数据集上进行测试,实验结果表明本文的方法在 C4.5 决策树上比其他的算法有更加好的划分精确度。

1　基于统计学法则的属性值划分方法

1.1　属性值划分方法描述

描述提出的决策树算法中基于统计学法则的连续属性值的新的数据划分方法在详细描述我们的算法之前,首先描述关于属性值数据划分的一些概念。

一个连续属性值的划分需要一个训练数据,它由 N 个样本(samples)组成,每个样本(samples)都属于 S 个类(class)中的一个。定义一个连续属性值的划分模型 P,P 可以将连续的属性值域划分到离散的 I 个区间,这些区间由一对数字来绑定,P 的描述如下:

$$P:\{[t_0,t_1],[t_1,t_2],\cdots,[t_{I-1},t_I]\} \tag{1}$$

式(1)中,t_0 是连续属性 P 的最小值,t_I 是 P 的最大值,划分模型 P 的值按升序进行排列。为达到属性值数据划分的目的,整个数据集都必须在一个连续的属性值中,这些数据集可以组成如表 1 所示的二维表。

2017 年 12 月 9 日收到　　国家自然科学基金(61640020)资助
第一作者简介:高洪涛(1965—),男,辽宁沈阳人, 硕士,副教授。研究方向:电子数据取证。E-mail:gaohongtao_2017@126.com。

表 1 连续的属性值

Table 1 Continuous attributes

区间	类的标签				行求和
	Class 1	Class 2	…	Class S	
$T_1:[t_0, t_1]$	N_{11}	N_{12}	…	N_{1S}	$N_{1.}$
$T_2:[t_1, t_2]$	N_{21}	N_{22}	…	N_{2S}	$N_{2.}$
⋮	⋮	⋮		⋮	⋮
$T_I:[t_{I-1}, t_I]$	N_{I1}	N_{I2}	…	N_{IS}	$N_{I.}$
列求和	$N_{.1}$	$N_{.2}$	…	$N_{.S}$	N(total)

表 1 中,I 表示行,S 表示列,每一行代表着初始的数据区间,每一列表示了一个不同的类(Class)。N_{ij} 表示 T_i 的 i 区间 j 类样本数,$N_{.j}$ 指 j 类的所有的样本数,$N_{i.}$ 指 T_i 的 i 区间内的所有的样本数。

正如所知道的,基于 Chi2 的数据划分算法都是基于 Bottom-up 自底向上的,建立在统计相互依赖的基础上的离散化方法,在此基础上,还有两类改进的 Chi2 算法:即改进的 Chi2 算法[6]和扩展的 Chi2 算法[7]。

改进的 Chi2 算法使用测量距离由 χ^2 修改为 χ^2_α,其中 α 为合并连续性区间的标准,这个合并标准可以用断点公式表示:

$$D = \chi^2_\alpha - \chi^2 \tag{2}$$

式(2)中,χ^2_α 的值由一个理想的参数 α 来决定,χ^2 的值可以解释为在目标类和划分属性值之间的相互依赖的假设的距离,可以用式(3)表示:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{S}\frac{(N_{ij} - E_{ij})^2}{N_{ij}} \tag{3}$$

式(3)中,$E_{ij} = \dfrac{N_{i.}N_{.j}}{N}$ 是 N_{ij} 期待的频率,并且有 $N = \sum_{i=1}^{I} N_{i.}$;$\dfrac{N_{.j}}{N}$ 是第 j 类样本数与所有的样本数的比值,E_{ij} 是在上述可能性的条件下,第 j 类样本数中第 $i \in P$ 区间模式的数量。因此,统计量 χ^2 显示了第 j 类样本数分配到相邻的区间的平均程度,χ^2 的值越小,那么类的分配就越接近,这样它也对于断点(cut point)来说越不重要,所以,在这个条件下的两个区间就必须合并。

考虑到断点 $D = x^2_\alpha - x^2$ 的重要性,因为断点是为了在确定的意义级别下加快增加系统中节点的统计性能,χ^2 值不仅和相邻的 2 个区间的数字的自由度有关系,而且和系统中的类的个数有关系。

这就意味着,系统中的类的数目越多,那么相邻的两个区间就有更加多的类的数目,那么相邻的两个区间的自由度就越大。因此 χ^2 的值的计算是不同的,根据属性值数据划分统计学准则,用式(3)进行计算:

$$Stacri = \chi^2_\alpha - \sqrt{\frac{2v}{S}}\chi^2 \tag{3}$$

实际上,这个重要统计学法则 *Stacri* 是根据 χ^2 来决定的,因为 $\sqrt{\dfrac{2v}{S}}$ 在 χ^2 的前面,即使按照比例减少 χ^2 的值,也不会减少最初的 α 值。这样也显示出了在区间合并的时候足够的因素,所以本文的新方法考虑了一个新的合并法则,即 $Stacri = \chi^2_\alpha - \sqrt{\dfrac{2v}{S}}\chi^2$ 作为基本的区间合并法则。

1.2 闭渐进项集

使用伪代码设计了一个连续属性值划分的启发式算法,该算法的目的是发现最优的数据划分,包括贪婪自底向上的数据划分,算法的详细步骤可以用下面的过程来表示。

```
1: Input: data set with N examples and S target classes
2: Output: partition scheme with a set of intervals for each continuous attribute
3: Initialize:
4: Set the significance level as α = 0.5 and Calculate the level of consistency of the original data
5: Sort attribute values in ascending order
6: for each attribute do
7:         Sort attribute values in ascending order
8:         Compute χ² value and corresponding threshold of each two adjacent intervals
9: end for
10: Stacri(data);
11: Phase 1: Merge intervals considering all the attributes
12: while α can be decreased do
13:     while Merge(data) do
14:         if the level of consistency is changed then
15:             SubStacri(att);
16:         else
17:             Goto line 22
18:         end if
19:     end while
20: α0 = α; decreasing the significance level αby one level
21: end while
22: Phase 2: Refine intervals considering single attribute according to parallel partition strategy:
23: Set SL[i] = α0
24: Do until no attribute can be merged
25: for each mergeable attribute do
26:     while SL[i] can be decreased do
27:         while Merge(att) do
28:             if the level of consistency is changed then
29:                 SubD(att);
30:             end if
31:         end while
32:         Decreasing the significance level SL[i] by one level
33:     end while
34: end for
```

上述算法中 Merge(data)考虑到了子空间中的

所有属性值的每一个区间对(intervals pairs)。拥有最大的 Stacri 值的两个相邻的区间可以合并。该函数的返回值包括 True 和 False。如果涉及满足对应的条件情况下的区间,就可以返回 True 或者 False。

Stacri(data)函数的功能是计算所有区间对的 *D* 值。*SubStacri* (att)函数的功能是计算那些需要更新的 *Stacri* 函数值。*Merge*(att)函数的功能与 Merge (data)类似,除了每个区间对(intervals pairs)在同一个属性值的情况下以外。

2　实验与性能分析

2.1　测试数据来源

测试提出的数据划分方法在真实情况下的性能,选择了 UCI 数据集。这些数据集中的数据具有种类多样;而且数据的容量大小也变化多样的特点。同时这些数据的属性值都是连续的,一致性的。这些数据包括了真实信息世界中各类信息,包括医疗信息数据,科学领域的数据,值得说明的是这些数据都曾经用来测试过以前的模式识别和机器学习的算法,可以说是数据挖掘决策树类的 BenchMark 的测试数据。这些属性值数据集的基本情况如表 2 所示。

表 2　破片侵彻试验结果

Table 2　Fragment penetration test results

数据集	连续属性数	离散的属性数	决策类别数	样本数
CPS	4	6	2	534
Iris	4	0	3	150
Auto	5	2	3	392
Breast	9	0	2	683
Ionosphere	32	2	2	351
Pima	8	0	2	768
Glass	9	0	6	214

2.2　实验与性能分析

数据划分方法与已经有的三种算法进行了比较:Mod-chi2 方法[6]是一个最常见的自底向上的算法;CAIM 方法[8]是一个新型的自顶向下的算法;MDLP 方法[11]是一个利用最小描述长度原理的信息熵方法。

在下面的实验中,每个连续的属性值都分别被上面的四个算法进行数据划分。在所有的数据集中使用了十折交叉验证,用来测试精度,十折交叉验证是常用的精度测试方法。

对上述的 4 个算法的测试结果如表 3,通过对表中的实验结果数据比较可以知道,平均起来,本文的基于统计学法则的数据划分算法具有最高的数据集划分精度,这个也证明了本文的方法优于常见的决策树数据划分方法,它是一个性能更高的划分模型。

表 3　算法的数据集划分精度比较

Table 3　The data set partition accuracy with different algorithm

数据集	划分精度/%			
	Mod-Chi2 算法	MDLP 算法	统计学法则的算法	CAIM 算法
CPS	61.7	64.66	64.87	65.85
Iris	94.67	93.76	94.67	92.67
Auto	80.75	81.52	85.64	75.83
Breast	96.77	96.47	97.5	96.77
Ionosphere	92.73	92.14	92.73	92.01
Pima	72.85	72.55	76.73	73.41
Glass	43.91	43.76	49.62	44.29

3　结论

提出了决策树算法中新型的连续属性值划分方法,它使用统计学法则来完成连续区间的合并。在此基础上,还描述了一个可以达到理想数据划分结果的启发式算法。通过 UCI 数据集中的大量种类繁杂的真实数据对算法进行了测试,实验结果表明本文的算法优于常见的连续属性值划分方法,划分精确度良好,可以改善 C4.5 决策树的分类学习精度。

参　考　文　献

1　Liu H, Hussain F, Tan C L, *et al*. Discretization: An enabling technique. Journal of Data Mining and Knowledge Discovery, 2002;(6)4: 393—423

2　Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous feature. Proceedings of 12th International Conference of Machine Learning. San Francisco: Morgan Kaufmann Publisshers, 1995:194—202

3　周世昊,倪衍森. 基于类-属性关联度的启发式离散化技术. 控制与决策, 2011;26(10):1504—1510
Zhou Shihao, Ni Yansen. Heuristic discretization technique based on the class-attribute interdependence. Control and Decision, 2011;26(10):1504—1510

4　陈爱萍,张光会. 一种连续属性值域划分的离散化新方法,计算机应用研究, 2012;29(4):1307—1310
Chen Aiping, Zhang Guanghui. Novel discretization method for value domain partition of continuous attributes. Application Research of Computers, 2012;29(4):1307—1310

5　侯居在,梁　莹,任长志. 多变量连续属性离散化方法. 模式识别与人工智能, 2011;24(6):792—797
Ren Juchi, Liang Ying, Ren Changzhi. A multivariate discretization method for continuous attributes. Pattern Recognition and Artificial Intelligence, 2011;24(6):792—797

6　Tay E H, Shen L. A modified chi2 algorithm for discretization. IEEE Transactions on Knowledge and Data Engineering, 2002;14(3):

666—670

7 Su C T, Hsu J H. An extended chi2 algorithm for discretization of real value attributes. IEEE Transactions on Knowledge and Data Engineering, 2005;17(3):437—441

8 Kurgan L A, Cios K J. CAIM discretization algorithm. IEEE Transactions on Knowledge and Data Engineering, 2004;16(2):145—153

9 Catlett J. On changing continuous attributes into ordered discrete attributes. Proceedings of Fifth European Working Session on Learning. Berlin: Springer-Verlag, 1991:164—177

10 Biba M, Esposito F, Ferilli S, *et al*. Unsupervised discretization using kernel density estimation. The Twentieth International Joint Conference on Artificial Intelligence. Hyderabad: AAAI Press, 2007:696-7101

11 Fayyad U, Irani K, Multi-interval discretization of continuous-valued attributes for classification learning. Proc of Thirteenth International Joint Conference on Artificial Intelligence. San Mateo: Morgan Kaufmann, 1993:1022—1027

12 Bondu A, Boulle M, Lemaire V, *et al*. A non-parametric semi-supervised discretization method. 2008 Eighth IEEE International Conference on Data Mining. Pisa:IEEE,2008:53—62

13 Armengol E, Garcia-Cerdana A. Refining discretizations of continuous-valued attributes. Modeling of Decisions of Artificial Intelligence Conference. Berlin: Springer, 2013:258—269

14 Salvador G, Julian L, Antonio S J, *et al*. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 2013; 25(4):734—750

Partition Approach for Continuous Attributes Based on Statistical Criterion

GAO Hong-tao [1], LU Wei[2], YANG Yu-wang[2]

(Department of Cyber Crime Investigation, Criminal Investigation Police University of China[1], Shenyang 110035, China;
School of Computer Science and Engineering, Nanjing University of Science and Technology[2], Nanjing 210094, China)

[**Abstract**] Many classification algorithms such as ID3/C4.5/C5.0 decision tree algorithms rely on discrete attributes and need to quantify continuous attributes into a finite number of intervals. A new data partition method for continuous attributes was presented. This approach used a statistical criterion to discover the accurate discrete intervals which was required to merge. In order to promote the classification performance of decision tree algorithm, a heuristic algorithm was also discussed to gain excellent the quantify results. A serials of simulation had been done using UCI data sets. The experiments results and performance analysis show approach is a good partition model, C4.5 decision tree classification algorithm can benefit a lot from our method.

[**Key words**] continuous attributes learning accuracy statistical criterion classification algorithms