文章编号：2095-2163(2021)04-0077-06 中图分类号：TP181 文献标志码：A

数据驱动的属性值分类法及其在属性泛化中的应用

钱卓昊
(西安石油大学 计算机学院，西安 710065)

摘 要：现实中基于树型层次结构的属性值分类是普遍存在的，反映这种树型层次结构的属性值分类法(Attribute Value Taxonomy，AVT)已被证明对数据的泛化上是有效的。部分数据集已具备相关专家提供的 AVT，但大多数数据集不具备人为提供的 AVT。为此，本文提出一种 VDM-AVT 学习器，即一种依据数据自动构造 AVT 的方法；为了评价所构造 AVT 的质量，基于 VDM-AVT 学习器提出了 VDM-AVT-AGR 模型。VDM-AVT 学习器基于 VDM 距离，利用层次聚类将属性值抽象为树型层次结构，VDM-AVT-AGR 模型利用 VDM-AVT 学习器得到的 AVT 对数据进行属性泛化约简。实验表明，利用 VDM-AVT-AGRD 模型处理后的数据集相比原始数据集具有更好的分类性能和泛化能力。由此也可以证明 VDM-AVT 学习器得出的 AVT 是有效的。

关键词：层次聚类；属性泛化约简；VDM；属性值分类法

Data-driven attribute value taxonomy and its application in attribute generalization

QIAN Zhuohao
(College of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 The classification of attribute value of tree hierarchy is commonly encountered. The Attribute Value Taxonomy (AVT) has been proved to be effective in data generalization. In some fields, the AVT has been designed artificially by experts. But most datasets do not have AVT provided by experts. Therefore, a VDM-AVT learner is proposed, which is a method to automatically construct AVT based on data. To evaluate the quality of the constructed AVT, the VDM-AVT-AGR model is proposed based on VDM-AVT learner. Based on the VDM distance, the VDM-AVT learner abstractions the attribute values into tree hierarchy using hierarchical clustering. The VDM-AVT-AGR model uses the AVT obtained from the data processing by the VDM-AVT learner to realize attribute generalization reduction. The experiments show that the data set processed by VDM-AVT-AGRD model has better classification performance and generalization ability than the original data set. It can also be proved that the AVT obtained by VDM-AVT learner is effective.

【Key words】 Hierarchical clustering; Attribute generalization; VDM; Attribute value taxonomy

0 引 言

现实生活中，属性值分类(AVT)又称层次属性值(Hierarchical Attribute Value)，是广泛存在的，如时间属性上日、月、季、年等具有层次特征的属性值[1]。可以利用概念层次将原始基础数据抽象到不同层次，实现数据泛化。同时，基于多层次(Multiple Levels)数据挖掘，可能会从较高层次数据中发现更普遍或更重要的知识，且获取的规则也更易于理解[2]。数据集中 AVT 树型结构可由相关领域专家提供，也可根据训练集自动构建而成。

具有层次结构的数据已被广泛应用，Han 等提出了一种利用概念分类法和自顶向下递进深化方法在不同层次上寻找概念之间的关联规则的算法 XLT2L1[3]；如 Hong 等基于粗糙集理论提出一种获取跨层次确定性规则和可能性规则的方法[4]；研究了具有层次结构的模糊粗糙集[5]；Feng 等利用层次结构提出一种自上向下的挖掘层次决策规则的方法[6]。

虽然 AVT 的有效性已被证明，但针对构建 AVT 的研究还比较少。涉及 AVT 时，大多是基于相关专家意见所构建的 AVT，这使得 AVT 具有主观成分，且在研究高维度数据时其准确度降低。在已有的从数据中构建 AVT 研究中，都是将 AVT 直接与分类模型综合在一起来处理数据，而没有进行属性泛化约简，这使得处理后的数据依旧可以进一步泛化。如 AVT-NBL 模型，在构建 AVT 时如何度量属性值间的相似关系目前也没有最佳标准，用 JS 散度来度

基金项目：国家自然科学基金(61976244)。

作者简介：钱卓昊(1995-)，男，硕士研究生，主要研究方向：粗糙集、机器学习。

收稿日期：2020-12-04

量[1]。为了研究 AVT 在离散属性中的应用,本文采用 VDM 距离来度量。

本文利用 VDM 度量样本属性值间的距离,进而利用层次聚类设计了一种依据数据自动构建 AVT 的 VDM-AVT 学习器。为了验证这种学习器的有效性,本文构建了 VDM-AVT-AGR 模型,该模型在处理数据集时,会基于 VDM-AVT 学习器对数据构建的 AVT 层次模型来进行属性泛化约简,使得数据集在属性个数和属性域两个层面实现降维。

1 基本概念

1.1 决策表

决策表(也可称为决策信息系统)是一个四元组 $S=(U,C\cup D,V,f)$,其中 U 是非空有限对象集,称为论域;$C\cup D$ 是非空有限属性集;C 为条件属性集;$D=\{d\}$ 为决策属性集,且 $C\cap D=\varnothing$;$V=\bigcup_{a\in C\cup D}V_a$,$V_a$ 为属性 a 的值域;$f:U\times A\rightarrow V$ 为信息函数[7]。

1.2 属性值分类法(Attribute value taxonomy)

VDM-AVT 学习器所构建的 AVT 中,V_a 是包含属性 a 的所有初始值的有限集,属性 a 的属性值分类 $AVT(a)$ 是基于 V_a 内元素间的相似性构建的一种树形概念层次结构。

$AVTs(C)=\{AVT(a_1),AVT(a_2),\ldots,AVT(a_m)\}$ 是关于 $C=\{a_1,a_2,\ldots,a_m\}$ 的属性值分类集合。$AVT(a)$ 中,叶节点相当于属性 a 在原始数据中的初始属性值。内部节点代表其相应子节点的泛化属性值,树上的每段弧代表了相邻且不同层次属性值之间的粗化或细化关系。令 $Leaf(a)$ 代表 $AVT(a)$ 的叶节点,$Root(a)$ 代表其根节点,$Node(a)$ 代表 $AVT(a)$ 的所有节点,内部节点(除叶节点以外的节点)相当于属性 a 的泛化值。$Child(v,a)$ 是属性值 v 在 $AVT(a)$ 中的所有子节点集合,$Depth(a)$ 是 $AVT(a)$ 中从根节点到叶节点的最大路径长度。如图 1 为本文利用 VDM-AVT 学习器处理 UCI 数据集 Car Evaluation 的 maint 属性得到的 $AVT(maint)$。

其中:

$Leaf(maint)=\{vhigh,high,med,low\}$;

$Root(maint)=\{vhigh+high+med+low\}$;

$Node(maint)=\{vhigh,high,med,low,vhigh+high,med+low,vhigh+high+med+low\}$;

$Child(med+low,maint)=\{med,low\}$;

$Depth(maint)=3$.

1.3 层次聚类

AVT 的构建可看作一个层次聚类的过程。层次聚类法(hierarchical methods)是递归地对数据对象进行合并或者分裂,直到某种终止条件满足为止。根据层次的分解的方式,具体又可分为"自底向上"和"自顶向下"两种方案。在"自底向上"方法当中,初始时将每个对象作为单独的一个聚类,相继地合并相互邻近的聚类,合并一次之后,聚类总数减一,直到所有的聚类合并成一个聚类或是满足一个终止条件[8-9]。

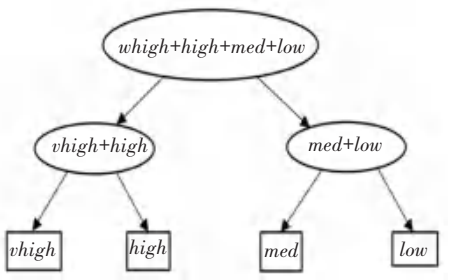


图 1 AVT(maint)

Fig. 1 AVT(maint)

VDM-AVT 学习器采用"自底向上"的层次聚类思路对属性不断进行抽象。先计算样本之间的距离,每次将距离最近的点合并到同一个类;再计算类与类之间的距离,将距离最近的类合并为一个大类;不停的合并,直到合成了一个类。最短距离法是将类与类的距离定义为类与类之间样本的最短距离[10]。而 VDM-AVT 学习器将 VDM 作为距离度量,同时每次选择距离最短的类进行合并。

1.4 VDM(Value Difference Metric)

现实生活中存在许多类似交通工具(火车,汽车)这样的无序属性,本文主要针对这类型数据进行研究,VDM 距离在处理离散属性时的有效性已经被广泛验证,比如 Zhang 等在处理异构数据的离散部分时使用 VDM 距离度量数据相似程度[11]。本文亦选择将 VDM 作为层次聚类的距离定义。

令 $m_{u,x}$ 表示属性 u 上取值为 x 的样本数;$m_{u,x,i}$ 表示在第 i 个样本簇中在属性 u 上取值为 x 的样本数;k 为样本簇数,则属性 u 上两个离散值 x 与 y 之间的 VDM 距离为式(1):

$$VDM_p(x,y)=\sum_{i=1}^{k}\left|\frac{m_{u,x,i}}{m_{u,x}}-\frac{m_{u,y,i}}{m_{u,y}}\right|^{p}. \quad (1)$$

其中,i 簇在决策表中对应决策属性的属性值,i 簇所包含的对象即是决策属性值为 i 的对象;k 为决策属性的属性值种类数;则可计算属性 u 上两个离散值 x 与 y 之间的 VDM 距离。

1.5 局部分割

对于属性 a 的属性值分类 $AVT(a)$，其局部分割 γ 定义为 $Node(a)$ 的子集，且 γ 满足以下性质：

(1)对于任意 $p \in Leaf(a)$，则 $p \in \gamma$，否则，存在 $q \in \gamma$，p 是 q 的祖先；

(2)对于 γ 中任意二个结点 p 和 q，p 既不是 q 的祖先，也不是 q 的后代[7]。

对于属性 a 中任何给定的一个局部分割 γ，其抽象属性值域形成一个较低层次局部分割中属性值的一个划分，也给出属性 a 的所有初始值的一个划分。局部分割将一个属性 a 的不同初始值抽象到不同层次，形成该属性在其概念层次 $AVT(a)$ 上不同层次的值域[12]。

1.6 属性泛化约简

属性约简可有效降低单尺度数据的维度，但其只能减少属性的数量，不适用于具有层次结构的属性。

属性泛化可以利用属性值分类法(AVTs)将原始属性的属性值转换为更粗的粒度。具体来讲，即在 AVT 中查找一个局部分割，使得在用这个分割所包含属性域来代替原始数据属性域后，数据集所包含的信息量并未改变。以信息量不变为前提，在 AVT 中查找尽可能粗的分割，就可以使数据集在分类能力不变弱的同时，属性泛化程度最大化。如基于香农熵提出了一种属性泛化约简算法 AGR-SCE[7]。

2 VDM-AVT 学习器及其评估

2.1 VDM-AVT 学习器实现方法

VDM-AVT 学习器对决策表的每个条件属性分别构造 AVT，最后得到一个由各个条件属性树形层次结构组成的 AVT 集合。单个属性 AVT 的构建可以看作是一个层次聚类过程，层次聚类的实现可分为“自底向上”和“自顶向下”两种方案，由于 VDM-AVT 学习器所研究的决策表一般是给出属性的初始属性值，也即属性值分类的叶子节点，AVT 的其它内部节点均为未知，因此选择“自底向上”方案作为 VDM-AVT 学习器构造 AVT 的方法。VDM-AVT 学习器在构造单个条件属性 AVT 时，先将该条件属性的原始值域中的元素作为叶子节点，然后分别计算各个叶子节点两两间的 VDM 距离，得出各叶子节点的概率分布，VDM 距离越小，表明对应的两个叶子节点的概率分布越相似，即对应叶子节点的相似性越大，因此将所有叶子节点中 VDM 距离最小的两个抽象为一个内部节点，这个内部节点是所抽象节点的父节点，再在新的节点集合中找出 VDM 距离最小的两个节点抽象为一个新的内部节点，重复此步骤直到所有叶子节点被抽象到只剩一个节点，此节点即为 AVT 的根节点，由此得出该条件属性的树形层次结构。具体构造过程见 VDM-AVT 学习器实现算法。

VDM-AVT 学习器实现算法：

Step 1　输入：决策表 $S=(U,C\cup D,V,f)$，其中，C 为条件属性集，D 为决策属性集；

Step 2　遍历所有条件属性，并对当前遍历到的属性 A_i 做 Step 3 至 Step 8 处理；

Step 3　计算该属性下各类属性值相互间的 VDM 距离，选取 VDM 距离最小的一对属性值 $v_i^{j_1}$ 和 $v_i^{j_2}$；

Step 4　将 $v_i^{j_1}$ 和 $v_i^{j_2}$ 合并为新的属性值 $v_i^{j_1+j_2}$，$v_i^{j_1+j_2}$ 为 $v_i^{j_1}$ 和 $v_i^{j_2}$ 的父节点；

Step 5　更新数据，将 $v_i^{j_1}$ 和 $v_i^{j_2}$ 对应对象在 A_i 属性的值改为 $v_i^{j_1+j_2}$，方便后续 VDM 计算；

Step 6　更新 A_i 属性值种类，删除 $v_i^{j_1}$ 和 $v_i^{j_2}$，并添加 $v_i^{j_1+j_2}$；

Step 7 如果此时 A_i 的属性值种类数量不为 1，从 Step 3 开始继续运行，否则运行 Step 8；

Step 8　得到 A_i 属性的 AVT 层次树 T_i；

Step 9　输出：$T=\{T_1,T_2,\ldots,T_n\}$。

2.2 VDM-AVT 学习器的评估

VDM-AVT 学习器的功能仅仅是构造出数据的树形层次结构，很难直接去评估所构建层次结构的好坏，但如果将 AVT 用于数据的泛化约简，所构造的 AVT 会直接影响泛化约简的结果，泛化约简结果的好坏也就反映了所构建 AVT 的有效性。也即如果 AVT 构造不合理，泛化约简的结果也很可能不合理，甚至导致出现数据不存在泛化约简的情况，用分类器对这样的泛化约简处理后的数据进行分类和规则提取，得出的分类准确率与用原始数据分类得到的分类结果相比是更低的，所提取的规则与从原始数据提取的规则相比也会更不合理或者所提取规则的数量不会减少；相反如果 AVT 构造合理，基于此进行泛化约简后的数据的分类性能相比原始数据会更好，所提取规则也具有更好的泛化性，具体量化表现为所提取规则的数量更少。为了评估 VDM-AVT 学习器的有效性，将 VDM-AVT 学习器所构建的 AVT 应用到泛化约简中，基于此提出 VDM-AVT-

AGR 模型。

AGR 代表属性泛化约简,相关研究已有很多,AGR-SCE 算法对数据集进行泛化约简,并用仿真实验验证了算法效果,证明了 AGR-SCE 算法可以实现数据条件属性的泛化约简,AVT 是基于相关领域知识构造的,具有主观性,且普适性不强[7]。VDM-AVT-AGR 模型将 AGR-SCE 算法中的 AVT 改为利用 VDM-AVT 学习器构造,使得算法可以应用于各种数据集而不需要学习相关领域知识。

3 实 验

通过对比利用 VDM-AVT-AGR 模型处理后的数据和原始数据的分类性能和所能提取的规则数,来验证 VDM-AVT-AGR 模型的有效性,进而证实 VDM-AVT 学习器的有效性。

从机器学习 UCI 数据库中选取了 4 个数据集进行研究,数据集中的非离散属性利用 spss 的分箱功能做离散化处理,数据集中包含缺失值的对象做删除处理。利用 VDM-AVT-AGR 模型对预处理后的数据集进行泛化约简得出泛化约简后的数据 AGRD (Attribute-generalization reduced data)。表 1~表 4 描述了将各数据输入 VDM-AVT 学习器后,利用各数据集生成的每个属性的 AVT 层次深度。

表 1 Breast Cancer Wisconsin (Original) 属性描述

Tab. 1 Description of the attribute of Breast Cancer Wisconsin (Original)

Attribute	Distinct Values	Depth of AVT
*Attribute*1	10	6
*Attribute*2	10	5
*Attribute*3	10	7
*Attribute*4	10	6
*Attribute*5	10	7
*Attribute*6	10	6
*Attribute*7	10	6
*Attribute*8	10	7
*Attribute*9	9	7

表 2 Car Evaluation 属性描述

Tab. 2 Description of the attribute of Car Evaluation

Attribute	Distinct Values	Depth of AVT
*Attribute*1	4	3
*Attribute*2	4	3
*Attribute*3	4	4
*Attribute*4	3	3
*Attribute*5	3	3
*Attribute*6	3	3

表 3 Iris 属性描述

Tab. 3 Description of the attribute of Iris

Attribute	Distinct Values	Depth of AVT
*Attribute*1	8	5
*Attribute*2	5	4
*Attribute*3	6	5
*Attribute*4	5	4

表 4 Wine 属性描述

Tab. 4 Description of the attribute of Wine

Attribute	Distinct Values	Depth of AVT
*Attribute*1	8	5
*Attribute*2	11	7
*Attribute*3	5	4
*Attribute*4	5	4
*Attribute*5	9	6
*Attribute*6	10	5
*Attribute*7	5	4
*Attribute*8	6	4
*Attribute*9	6	6
*Attribute*10	6	6
*Attribute*11	7	7
*Attribute*12	3	3
*Attribute*13	8	5

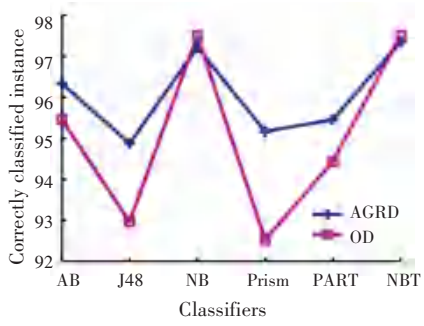
实验中,利用 Weka 软件,使用其中 6 种分类模型构建方法分别对各个数据集的 AGRD 和原始数据集 OD(the Original Data)进行分类,6 种模型分别是:朴素贝叶斯分类器(NB)、增强算法(AB)、两种决策树(J48 和 NBT)和两种基于规则的分类器(PART 和 Prism),所有分类器都使用 Weka 默认参数进行训练。最后通过十字交叉验证,即将实验数据随机分成 10 组,依次将其中的 9 组作为训练集,剩余的 1 组作为测试集,最后将 10 次实验结果的均值作为最终的验证结果,获得分类精度。同时利用 J48、NBT、PART、Prism4 种模型得出所提取的规则数目。

分类准确率反映了正确分类的实例数量占总实例数量的百分比,准确率越大分类性能越好。图 2 描述了各数据集在几种分类器下 AGRD 和 OD 的分类准确率比较,从中可以得出,AGRD 的分类性能是高于 OD 的。

规则数量决定了模型的复杂度,图 3 描述了各数据集分别在几种模型下可提取出的规则数量,可以看出从 AGRD 中提取的规则数量总体上明显少于 OD,同时 AGRD 的分类性能更好,因此其提取的

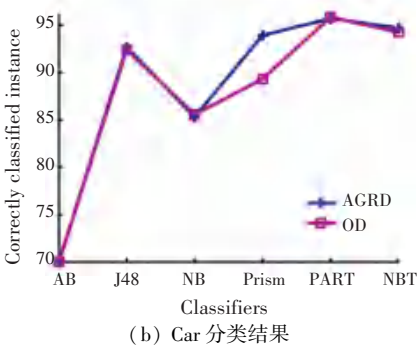
规则可靠性也更大。

通过以上实验分析，可以看出将 VDM－AVT－AGR 模型用于层次泛化约简可以有效提高数据分类性能，可以使数据在属性数量和属性值数量方面都变得更加简洁，从而实现更好的分类性能，提取出更可靠的分类规则。



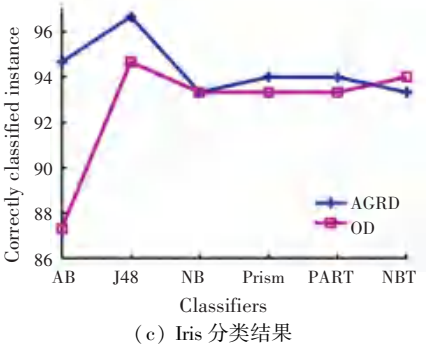
（a）Breast Cancer 分类结果

（a）The classification results of Breast Cancer



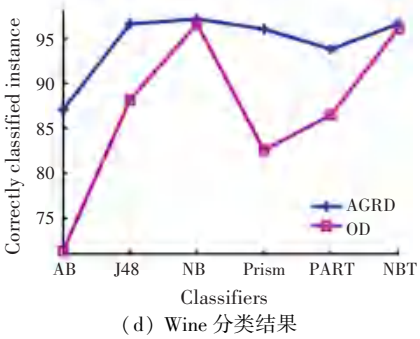
（b）Car 分类结果

（b）The classification results of Car



（c）Iris 分类结果

（c）The classification results of Iris

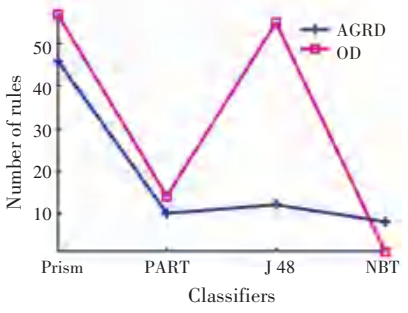


（d）Wine 分类结果

（d）The classification results of Wine

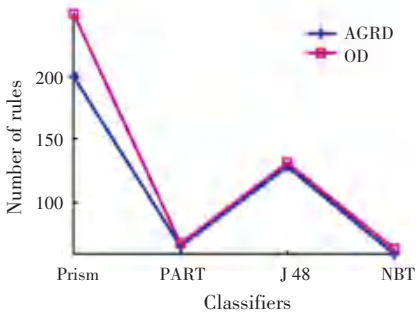
图 2　AGRD 与 OD 分类准确率对比

Fig. 2　Comparison of classification accuracy between AGRD and OD



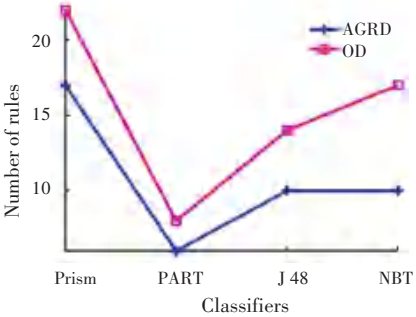
（a）Breast Cancer 规则提取结果

（a）The rule extraction results of Breast Cancer



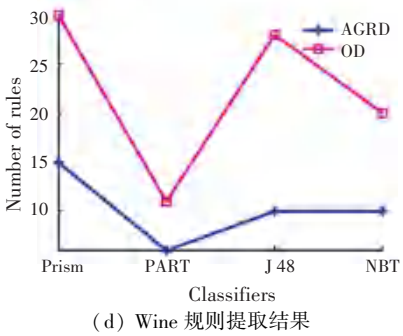
（b）Car 规则提取结果

（b）The rule extraction results of Car



（c）Iris 规则提取结果

（c）The rule extraction results of Iris



（d）Wine 规则提取结果

（d）The rule extraction results of Wine

图 3　AGRD 与 OD 提取规则数对比

Fig. 3　Comparison of extraction rule number between AGRD and OD

（下转第 87 页）