

Study of Attribute-Oriented Induction Based on Variance Analysis

LI Dongping

Kunming University, Kunming, Yunnan, China

e-mail: R3mj8@21cn.com, e-mail yfldp@sina.com

Abstract: Attribute-Oriented Induction is the most effective and convenient method of data mining to extract characterization rules of data. This paper introduces a new data mining method combining variance analysis and Attribute-Oriented Induction to solve the problem of choosing the class threshold in data generalization and the best characterization rules in data induction. This method can be more extensively applied in practice. The experimental results demonstrate the effectiveness and feasibility of the algorithm of Attribute-Oriented Induction based on variance analysis.

Keywords: attribute-oriented induction; variance analysis; characterization rules; data generalization

基于方差分析的 AOI 研究

李冬萍

昆明学院, 昆明, 中国, 650031

e-mail: R3mj8@21cn.com, yfldp@sina.com

【摘要】面向属性归纳 (AOI) 是提取数据特征规则最有效、方便的一种数据挖掘方法。文中将方差分析方法和面向属性归纳的方法结合起来,形成了一种可广泛应用的数据挖掘方法,解决了数据概化过程中阈值的选取和数据归纳中最优特征规则选取的问题,并用实验展示了基于方差的面向属性归纳算法有效性及可行性。

【关键词】面向属性归纳; 方差分析; 特征规则; 数据概化

1 引言

数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘获得的规则包括描述型规则和预测型规则,特征化规则是一种典型的描述性规则,数据概化是描述型规则挖掘方法的一种主要手段,而面向属性的归纳方法则是实现数据概化及生成特征化规则的一种重要方法。

本文从介绍面向属性归纳的方法入手,分析出面向属性归纳算法的不足,提出了一种新的属性阈值选择和最优特征选择的算法,即基于方差分析的面向属性归纳的算法,并将此新算法具体应用到昆明某汽车市场今年上半年购买不同种类家庭小汽车的消费者数据模型中,取得了较好的实际应用效果。

2 面向属性的数据归纳

面向属性的数据归纳 (Attribute-Oriented Induction, 简称为 AOI), 是一种面向关系数据库查询的、基于概化的、联机的数据分析处理技术,是一种用于数据库的知识发现^[1]。其主要思想是:使用关系数据库查询收集任务相关的数据,通过考察任务相关数据中每个属性的不同值的个数,进行概化;接着,使用聚集函数作用于概化后的数据来合并相等的广义元组,累计它们对应的计数值;然后,将经过数据压缩后得到的广义关系映射成不同的形式 (如图表或规则) 提供给用户。在现实中,通常让决策者感兴趣的规律一般只有在较高的概念层面上才能发现,并且希望结果表达得简洁一些。所以, AOI 可以有多种不同的应用,其中之一是它能够在一个数据挖掘过程中,被运用到处理初始数据之间的关系,效果是简化分类所要处理的事件空间以及得到一

些较高概念层上的广义关系。

面向属性归纳的基本操作是数据概化，它可用属性删除和属性概化来实现^[2,3]。

(1) 属性删除：如果初始关系的某个属性包含大量的不同值，那么符合下列两种情况之一时该属性就应该被删除。其一，在该属性上没有概化操作；其二，它的较高层概念已经有其他属性表示。

(2) 属性概化。如果某个属性包含有大量的不同值，且在该属性上有概化操作符，则运用该操作符对此属性进行概化。这样做的目的是使得到的规则涵盖更多的原始数据元组，从而概化了它所表示的概念。

属性删除和属性概化是面向属性归纳方法之所以能具有效率高、表达结果简洁明了等特征的主要操作，一个属性是否要删除、其概化程度的高低取决于属性本身或应用。如果属性概化得过高，会导致过分概化，产生的规则可能没有多少信息；如果属性不能在足够高的层次概化，会导致概化不足，最终得不到最优信息。因此需要对属性概化过程进行有效的控制。常用的控制方法有两种：属性概化阈值控制（ α ）和概化关系阈值控制（ β ）。属性概化阈值的取值一般为 2—8；概化关系阈值的取值一般为 10—30。这两种阈值一般是由用户或专家设定，如果属性的不同值个数大于属性概化阈值，则应当进行进一步的属性删除或属性概化，如此反复，直至不同属性值数目小于等于属性阈值；如果不同元组的个数大于概化关系阈值控制，则应当进行进一步概化。在实际操作中可以先使用属性概化阈值控制技术概化每个属性，再使用概化关系阈值控制进一步压缩概化关系。

经过 AOI 的数据概化后，初始数据集中会出现很多内容相同的数据行，此时合并内容相同的数据行，并累计这个唯一数据行对应的 count 属性值。关系数据库经过 AOI 后，不同的记录数目会大量减少，我们可依据这些记录的广义关系生成概化结果，即特征规则。此概化结果可用表、图的形式表示定性的描述，也可用逻辑规则来表示定量的描述，通常每个泛化后的数据行代表（概念描述）规则中的一个析取项。由于一个大型数据库中的数据通常具有多种不同的分布；因此一个泛化后的数据行不可能覆盖或表达所有（100%）初始数据集中的数据行。因此生成定量信息时，我们把满足规则条件左边（自然也满足规则右边）的数据行数目与初始数据集中总行数之比，作为所获概念描述规则的一个度量客观价值的重要参量，称作 t-weight 权值。公式为：

$$t_weight = count(q_a) / \sum_{i=1}^N count(q_i) \quad (1)$$

其中：n 是泛化后目标集合中的数据行个数； $q_1, \dots, q_n$ 是泛化后目标集合中数据行，显然 q_a 在 $q_1, \dots, q_n$ 中，而 t-weight 取值在 [0, 1] 区间中。

那么一个定量概念描述规则可以用：(1) 带有 t-weight 的各析取项所组成的逻辑规则形式所表示；(2) 关系表或组合表形式来表示；其中 count 值换成对应的 t-weight。定量的逻辑规则形式如：

$$\forall X, \quad target_class(x) \Rightarrow condition1(X)[t : w_1] \vee \dots \vee conditionn(x)[t : w_n]$$

上述规则表明：若 X 在目标集合中，那么有 W_1 概率 X 满足条件要求；这里 W_i 就是条件或析取项 i 的 t-weight，i 为 {1, 2, 3, ..., n}。

在进行客观评估时，用户可根据需要设定一个阈值，当一个泛化后数据行的 t-weight 小于该阈值时，就可认为该数据行仅代表了一小部分（可忽略）的数据库内容，继而被认为是无意义且可被忽略。但忽略这部分数据行并不意味着它们也应该从所获得的（中间）挖掘结果中消除掉，这主要考虑到这些数据行稍后可能会在更进一步的（用户）数据探索过程中使用到。

综上所述，我们发现，特征规则提取是通过将定量信息与特征规则相结合定量地反映了概念的主要特征和次要特征的，挖掘特征规则是面向属性的归纳学习算法的最终实现目标。但我们好发现，在面向属性归纳算法的核心步骤数据概化中，阈值大小的选取是关键，而不同的专家、用户采用的属性概化阈值和概化关系阈值不同，这时，在选择哪个属性上进行再次提升时，就没有固定的标准，从而导致在不同的数据挖掘过程中造成概化的结果不同，那么，同一个概念就会有不同的特征描述产生^[4,5]，这种现象必然会降低特征规则对决策的参考价值。针对这一问题，本文提出了一个衡量阈值大小和特征规则有效性的计算指标：方差化分析，来解决这个问题。从而使定量特征规则在有效性上达到最优，避免了在不同的挖掘过程中同一概念有不同的特征规则的现象。

3 改进的面向属性的归纳（R-AOI）

3.1 方差分析法

方差是各个数据与平均数之差的平方的平均数，也就是目标数据和中心偏离的程度，用来衡量一批数据波动的大小（即这批数据偏离平均数的大小），方差越大，同一批数据波动越大。

方差分析是在使用属性泛化阈值 α 控制来泛化每

个属性后，泛化关系后内容不相同的行数大于泛化关系阈值 β ，然后选择一个属性进行攀升，同时计算方差，当攀升后，内容不相同的行数目 $\leq \beta$ ，且方差最大的规则，就是最优的规则。

定性特征规则反映的是属于某概念的大部数据所共同具有的特征。而定量特征规则是覆盖了所有属于该概念的数据，它通过定量信息，反映了概念的主要特征和次要特征，因此能否将概念的主要特征与次要特征区分开是衡量定量特征规则有效性的重要标志，那么，我们设有定量特征为：

$$\forall X, \text{target_class}(x) \Rightarrow \text{condition1}(X)[t : w_1] \vee \dots \vee \text{condition}^\beta(x)[t : w_B]$$

其中 β 为用户定义的泛化关系阈值，用户希望生成 β 类，若 X 在目标集合中，那么有 w_1 概率 X 满足条件要求，这里 w_i 就是条件或析取项 i 的 t -weight， i 为 $\{1, 2, 3, \dots, \beta\}$ 。那么：设指标 V 是对定量特征规则有效性进行度量的数据指标，利用方差公式得：

$$V = \frac{\sum_{i=1}^{\beta} (w_i - \bar{w})^2}{\beta - 1} \quad (2)$$

其中， β 为泛化关系阈值， w_i 是第 i 项特征析取项的 t -weight 值， i 为 $\{1, 2, 3, \dots, \beta\}$ ，

$$\bar{w} = \frac{\sum_{i=1}^{\beta} w_i}{\beta} \quad \text{因为} \quad \sum_{i=1}^{\beta} w_i = 1, \quad \text{所以} \quad \bar{w} = \frac{1}{\beta}. \quad (3)$$

$\bar{w}$ 为 β 中每一类（样本）的平均数。

3.2 改进的面向属性归纳的算法实现步骤 (R-AOI)

面向属性的归纳算法挖掘特征规则是针对每一个概念进行挖掘的，其任务就是从数据中挖掘出每一类数据的特征规则，最后，我们再用方差分析的方法选择最佳的特征规则作为广义关系映射成图表或规则提供给用户。具体操作如下：

- (1) 从关系数据库中查询并收集与任务相关的数据；
- (2) 数据预处理：包括把连续数据转换成离散值；
- (3) 采用不同的阈值调用属性归纳算法模块，生成数据集；
- (4) 生成最佳规则：利用步骤三模块生成初始特征规则，计算有效指标 V 。最后提取最佳规则；
- (5) 输出规律性描述（图表或规则）。

4 实验

我采用的实验数据是昆明某汽车市场今年上半年购买不同种类家庭小汽车的消费者数据，样本量为 60000 条记录。决策属性为“buy_car”，条件属性为：年薪、年龄、是否拥有自己的住房。我选择不同的控制阈值进行属性概化，经过数据归纳后，我们得到表达式为 buy_car = “Camry 240G”的 4 类特征描述：

第一类：buy_car = “Camry 240G” $\rightarrow$ age = “40-50” and salary = “100,000 — 200,000” and own of house = “yes” [80%] or age = “30-40” and salary = “100,000 — 200,000” and own of house = “yes” [12%] or age = “50-60” and salary = “50,000 — 100,000” and own of house = “yes” [7.5%] or age = “20-30” and salary = “50,000 — 100,000” and own of house = “no” [0.5%]。

第二类：buy_car = “Camry 240G” $\rightarrow$ age = “40-50” and salary = “150,000 — 200,000” and own of house = “yes” [70%] or age = “30-40” and salary = “100,000-150,000” and own of house = “yes” [20%] or age = “50-60” and salary = “100,000 — 200,000” and own of house = “yes” [7%] or age = “20-30” and salary = “50,000 — 100,000” and own of house = “no” [3%]。

第三类：buy_car = “Camry 240G” $\rightarrow$ age = “40-60” and salary = “150,000 — 200,000” and own of house = “yes” [80%] or age = “20-40” and salary = “100,000 — 150,000” and own of house = “yes” [15%] or age = “40-60” and salary = “50,000 — 100,000” and own of house = “yes” [3%] or age = “20-40” and salary = “50,000 — 100,000” and own of house = “no” [2%]。

第四类：buy_car = “Camry 240G” $\rightarrow$ age = “20-60” and salary = “150,000 — 200,000” and own of house = “yes” [25%] or age = “20-60” and salary = “100,000-150,000” and own of house = “yes” [25%] or age = “20-60” and salary = “100,000 — 200,000” and own of house = “yes” [25%] or age = “20-60” and salary = “50,000 — 100,000” and own of house = “no” [25%]。

取出各类第 i 项特征析取项的 t -weight 值为：

第一类：W1 = 0.8 W2 = 0.12 W3 = 0.075 W4 = 0.05
 第二类：W1 = 0.7 W2 = 0.2 W3 = 0.07 W4 = 0.03
 第三类：W1 = 0.8 W2 = 0.15 W3 = 0.03 W4 = 0.02
 第四类：W1 = 0.25 W2 = 0.25 W3 = 0.25 W4 = 0.25

利用定义(2) (3) 的公式对定量特征规则有效性进行度量的数据指标 V 进行计算，得到以上四类的方差为： $V1=0.136683$, $V2=0.095267$, $V3=0.137933$, $V4=0$ ，可

以看出，第三类的方差为 $V_3=0.137933$ 最大，而当方差为0时，对决策没有参考价。我们知道，方差越大，其主要特征和次要特征区别越明显，对决策的参考价值越好。所以第三类是最优的一种归纳关系。即购买Camry 240G汽车的主要客户是年龄在“40-60”、年薪在“150,000—200,000”且拥有自己的住房的人群，这类客户开发潜力最大；而年薪在100,000以下的客户购买Camry 240G汽车较少，开发潜力不大。那么利用此特征规则，决策者便可以作出正确的销售和广告宣传决策了。

5 结论

本文把方差分析和面向属性的数据归纳算法结合起来，用于客户购买能力的分析中，获得了最优的特征化规则，验证了R-AOI算法的正确性和可行性。实验证明，基于方差分析的面向属性归纳算法可以满足特征分类挖掘的要求，它具有高效、直观、类别特征比较明显

等特点。

References (参考文献)

- [1] Han J, Cai Y, Cercone N. Knowledge discovery in databases, an attribute-oriented approach[C]. In Proc. of the 18th VLDB Conference. Vancouver, British Columbia, Canada. 1992. P547 - P559.
- [2] Jiawei Han, Micheline Kamber, data mining concepts and techniques[M], Second Edition, Beijing, Machinery Industry Press, 2007 P128-P140.
- [3] Jiawei Han, Micheline Kamber 著，数据挖掘：概念与技术[M]，第2版，北京，机械工业出版社，2007 P128-P140.
- [4] YUE Yi, ZHENG Cheng, WU Guo-don, Application of Attribute Reduction in Inductive Learning-Based Intrusion Detection[J], Computer Technology And Development, Oct. 2007, Vol.17 No.10 P250-P254.
- [5] LIU Qi-hong, TANG Chang-jie, LI Chuan, LIU Qi-wei, ZENG Tao, JIANG Yong-guang, Traditional Chinese Medicine prescription mining based on attribute-oriented relevancy induction[J], Computer Applications, 2007 (27) 2, P449—452. 刘齐宏, 唐常杰, 李川, 刘齐巍, 曾涛, 蒋永光, 基于属性归纳的中药方剂数据挖掘[J], 计算机应用, 2007 (27) 2, P449—P452.
- [6] KDDCup99 [EB/OL]. 1999. <http://kdd.ics.uci.edu/da-tabases/kddcup99/kddcup99.html>, KDDCup.