

基于 FP-growth 算法的高维混合属性数据挖掘方法

梁树杰[†]

(广东茂名幼儿师范专科学校, 广东 茂名 525200)

摘 要: 常规高维混合属性数据挖掘方法多采用云平台技术, 无法完整保留数据的结构相似性, 使得数据挖掘效率较低。为此, 提出了基于 FP-growth 算法的高维混合属性数据挖掘方法。为了改善数据质量, 根据高维混合属性数据在数据库中的存储结构, 采用了一种固定算法实现数据去噪, 并依据数据类型计算分类型和数值型相似度, 结合 FP-growth 算法对频繁项样本分支进行筛选生成项表头, 保证数据结构相似性的完整性, 通过搜索项表头输出有效关联规则, 实现数据挖掘过程。实验结果表明, 所提方法具有较高的挖掘效率。

关键词: 数据挖掘; FP-growth 算法; 固定算法; 高维混合属性

中图分类号: TH17

文献标识码: A

A High Dimensional Mixed Attribute Data Mining Method Based on FP-growth Algorithm

LIANG Shujie[†]

(Guangdong Preschool Normal College in Maoming, Maoming, Guangdong 525200, China)

Abstract: Conventional high-dimensional mixed attribute data mining methods mostly use cloud platform technology, which can not completely preserve the structural similarity of data, making the efficiency of data mining low. For this reason, a high dimensional mixed attribute data mining method based on FP-growth algorithm is proposed. In order to improve the data quality, according to the storage structure of high-dimensional mixed attribute data in the database, a fixed algorithm is adopted to de-noise the data, and the classification and numerical similarity are calculated according to the data type, and the FP-growth algorithm is combined to filter the frequent item sample branches to generate the item header to ensure the integrity of the data structure similarity, and the data mining process is realized by outputting effective association rules from the search item header. Experimental results show that the proposed method has high mining efficiency.

Key words: data mining; FP-growth algorithm; fixed algorithm; high-dimensional mixed attributes

互联网环境中, 高维混合属性数据具有特征参数较多、属性复杂等特点, 为提高网络资源的利用率, 需要采取一些技术手段对数据进行挖掘, 扩大对数据的利用范围。如文献[1]利用差分进化算法对数据进行挖掘, 该方法通过采集数据库中的原始数据, 根据其结构类型建立采集模型, 并通过设定

模型的聚类中心, 采用差分进化算法计算数据的聚类数量, 根据样本结构对数据进行聚类分析, 最终实现数据挖掘过程, 该方法的挖掘稳定性较强, 但对于高维混合属性的大数据而言, 其灵活性较差, 实用性比较受限, 且挖掘准确性还需提高; 文献[2]利用物联网边缘计算对数据挖掘展开研究, 此方法

收稿日期: 2023-01-18

基金项目: 广东省教育厅资助项目(2022WTSCX302); 广东茂名幼儿师范专科学校 2022 年度教育科学“十四五”规划资助项目(2022GMYSKT08)

作者简介: 梁树杰(1981—), 男, 广东高州人, 硕士, 副教授, 研究方向: 数据挖掘技术, 信息安全技术等。

[†] 通信联系人, E-mail: tong06938940@163.com

通过对数据源进行分布式计算,求得边缘侧数据的调度系数,并利用贝叶斯关联分析搭建挖掘技术模型,以此完成数据挖掘任务,但该方法面对多层次的高维混合属性数据时,适应度和抗干扰性较差。

根据对以上高维混合属性数据挖掘方法的分析,提出了利用 FP-growth 算法对高维混合属性数据进行挖掘,从多组数据的相似度来提取数据集的频繁项集,并输出有效强关联规则,进而达到对数据可靠性挖掘的目的。

1 高维混合属性数据挖掘方法设计

1.1 高维混合属性数据去噪

高维混合属性数据在数据库中的存储结构为动态重组,数据样本本位置多分布于多维分块区域,其中包括大量的正样本数据、负样本数据及存在噪声干扰的多源数据。为稳定挖掘过程,提高挖掘质量,在对数据进行挖掘之前,需要执行去噪处理^[3]。本文采用固定算法实现数据去噪。

非高斯型度量函数如下:

$$J(y) = [E\{G(y)\} - E\{G(v)\}]^2 \quad (1)$$

上式中, $E(\cdot)$ 表示数据期望值; $G(\cdot)$ 表示数据的非线性函数; y 表示数据属性部分权重; v 表示绝对偏差。

为了快速提取到数据集集中的多个独立分值得,利用公式(2),将公式(1)的计算结果最大化^[4]:

$$J_G(w) = [E\{G(w^T p)\} - E\{J(y)\}]^2 \quad (2)$$

上式中, w 表示数据多维变量; p 表示高斯变量的协方差矩阵。

对度量函数最大化的过程实质上是对 $E\{G(w^T p)\}$ 的改进过程,当 $E\{G(w^T p)\} = \|w\|^2 = 1$ 时,可利用下式对最大化结果进一步优化,即:

$$w_{k+1} = \frac{w_k - [E\{G(w^T p)\} - \beta w_k]}{J_G(w) - \beta} \quad (3)$$

上式中, β 表示数据信息熵; k 表示迭代次数。

为使得数据去噪结果更加准确,当完成一次优化过程后,可采用标准化计算公式对优化结果 w_{k+1} 进行处理:

$$w_{k+1} / = \|w_{k+1}\| \quad (4)$$

接下来,将上式与式(3)结合,进行累加计算,得到 w_{k+1} 多次迭代之后的结果,公式如下:

$$r_k = E\{pG(w^T p)\} w_{k+1} \quad (5)$$

上式中, r_k 表示数据的独立分量。

式(5)表示初始权值矢量 w_0 为随机值,迭代次

数为 0 时对权值矢量 w_{k+1} 进行更新的结果。

利用上式对高维混合属性数据的估算分量进行提取,重复该算法以分离数据组件^[5]。为了保证提取出的数据组件之间内部不存在关联性,是相互独立的,需要每次提取出分量之后,在其余迭代数据里将此分量剔除,循环以上操作,即可提取出原始空间中所有数据的分量,完成去噪过程。

通过利用固定算法最大化数据的非高斯度量函数,并优化数据期望值,获取高维混合属性数据的独立分量,进而实现对数据的去噪处理,提高了数据库中数据的初始质量^[6],为计算数据相似度提供了便利条件。

1.2 高维混合属性数据相似度计算

高维混合属性数据在数据库中的属性通常表现为 4 种,包括方向、长度、参量和标题^[7]。计算数据相似度是进行数据挖掘的必要基础,因此,首先对多属性数据的相似度进行计算,以获取最终挖掘结果。

为了缓解网络系统中数据库的内存压力,提升外界算法的稳定性,基于数据去噪结果,利用空间维度变换原理,对数据进行矩阵变换。假设 D_j 表示数据库中第 j 列的结构矩阵, d_{ji} ($i = 1, 2, \dots, m$) 为其中的数据对象,表示第 i 列、第 j 行的数据。如果矩阵中每行共有 m 个高维混合属性数据,则 D_j 可表示为:

$$D_j = (d_{j1}, d_{j2}, \dots, d_{jm}) \quad (6)$$

如果结构矩阵中的数据对象数量为 n ,则使用矩阵的转置变换来表示数据库中高维混合属性数据的总矩阵 D ,如下所示:

$$D = (D_1, D_2, \dots, D_n)^T \quad (7)$$

现在,挖掘总矩阵 D 中集合 X 的高维混合属性数据 x ,并将其单个属性的关联度设置为 $\text{sim}(X, Y)$ (对应的集合由 Y 和 X 表示)。挖掘出的样本数据表示为 s ,则 s 应符合下列条件:

$$s = \text{sim}(X, Y) [\text{Freq}(x, X) - \text{Freq}(y, Y)] / D \quad (8)$$

上式中, Freq 表示条件发生的概率。

高维混合属性数据的尺度特征包括分类型和数值型两种^[8],假设数据集 X 中存在两条相互独立的数据 $X_i(x_{i1}, x_{i2}, \dots, x_{iq})$ 与 $X_j(x_{j1}, x_{j2}, \dots, x_{jq})$,那么两个数据间的距离由下式计算:

$$|X_i, X_j| = s \cdot \sum_{k=1}^q (x_{ik} - x_{jk})^2 \quad (9)$$

上式中, q 表示数据集 X 的维数; s 表示条件样本数据; x_{ik} 、 x_{jk} 分别表示 X_i 与 X_j 的第 k 维参

考样本。

数据 X_i 与簇 U_j 的相似度界定公式为:

$$|X_i, U_j| = |X_i, X_j| \sum_{k=1}^q (x_{ik} - U_{jk})^2 \quad (10)$$

上式中, U_j 表示数据簇的取值频度向量; U_{jk} 表示 k 维取值。数据 X_i 与参考样本数据之间的距离由下式计算:

$$H(v) = |X_i, U_j| \left(\frac{r_k}{2\sigma^2} \right) \quad (11)$$

上式中, r_k 表示数据的独立分量; σ 表示候选信息目标参数。

对于分类数据相似性的测量形式,如果两个分类数据 $X_i(x_{i1}, x_{i2}, \dots, x_{iq})$ 与 $X_j(x_{j1}, x_{j2}, \dots, x_{jq})$ 能够同时存在于一个数据矩阵中,则这两个数据间的相似性计算公式为:

$$S(X_i, X_j) = \frac{\sum_{k=1}^q \delta_{ij}(x_{ik}, x_{jk})}{H(v)} \quad (12)$$

上式中, δ_{ij} 表示高维混合属性数据的特征、位置、方向和长度属性同时出现的概率,计算方法为:

$$\delta(x_{1i}, x_{2i}) = \begin{cases} 0, & x_{1i} = x_{2i} \\ 1, & x_{1i} \neq x_{2i} \end{cases} \quad (13)$$

如果分类数据 X_i 和分类聚类 A 之间的相似度被认为是数据和其他数据之间的平均相似度^[9],则相似度的定义表达为:

$$M(X_i, A) = \frac{\|S(X_i, X_j)\|}{A \times q} \quad (14)$$

通过扩展上述定义,可以获得两种类型数据的聚类相似性,分别设置 A 、 B 是此类型数据的簇,则相似性界限如下:

$$N(A, B) = \frac{\|M(X_i, A)\|}{A \times B \times q} \quad (15)$$

将聚类界限作为数据相似度度量的提取标准^[10],并由以下方程描述:

$$l(x_i, U_j) = \frac{\|N(A, B)\|}{l_n + w_i l_c} \quad (16)$$

上式中, U_j 表示数据簇的取值频度向量; w_i 表示第 i 个数据对象的离散点代表度; l_n 、 l_c 分别表示 n 、 c 类数据属性类型的相似度:

$$l_n = l(x_i, U_j) \sum (x_{ik} - U_{jk})^2 \quad (17)$$

$$l_c = l(x_i, U_j) \sum \sigma (x_{ik} - x_{jk})^2 \quad (18)$$

以高维混合属性数据的属性分类为基础进行分析与计算,得到高维混合属性数据的数值型与分类型关联相似度^[11],便于实现对数据的挖掘。

1.3 利用 FP-growth 算法实现数据关联挖掘

FP-growth 算法是对传统数据挖掘方法的一

种改进算法,其最大的特点是:在已知高维混合属性数据相似度的情况下,无须再产生候选项集,而是将数据对象存储于 FP-树,再通过 FP-树直接得到频繁项集,实现过程中只对数据库遍历连词即可,极大地缩短了挖掘时间^[12]。

FP-growth 算法的实现过程主要分为两个步骤,第一个步骤的目标是构建一棵 FP-树,第二个步骤的目标是基于 FP-树,对数据对象进行关联规则挖掘。构建 FP-树是该算法执行过程中的关键环节,通过判断 FP-树除根节点以外的分支是否包含带有分类属性和数值属性的频繁项数据,来构成扩展树,若包含,则保留对应的节点;若不包含,则修剪该节点,继续向前扩展^[13]。当初次遍历完树中的所有节点与分支后,检查频繁项前一节点和后一节点中是否包含规则项,若包含,则保留该项集;若不包含,则生成相应的子节点。根据以上优化增长过程,得到基于改进的 FP 增长算法的高维混合属性数据挖掘的具体步骤如下。

(1)在增长树节点输入数据样本集,并利用下式计算最低支持度与最低置信度。

$$\min \gamma(X \rightarrow Y) = l_n \cdot \frac{\text{num}(XY)}{\text{num}(\text{total})} \quad (19)$$

$$\min \lambda(X \rightarrow Y) = l_c \cdot \frac{\text{num}(XY)}{\text{num}(X)} \quad (20)$$

其中, $\text{num}(\text{total})$ 表示数据参考集; $\text{num}(XY)$ 表示离群参考对象; $\text{num}(X)$ 表示样本点 X 到参考样本的平均距离。

(2)利用参考数据集扫描输入数据,根据最低支持度公式搜索支持度大于计算结果的频繁集,构建 FP 树和项目标题,并按照数值大小进行降序排列^[14]。

(3)对第一次扫描得到的频繁集进行第二次扫描,并根据支持度生成相应的数据节点。

根据样本分支是否包含高维混合属性数据的频繁项来过滤样本分支。如果是,则保留它们;否则,修剪分支,再次扫描,并重新生成项目标题和 FP 树。

(4)以项表头最后一个头指针为遍历起点,对表头中的所有数据项进行随机遍历,得到带有递归项的频繁项。

(5)判断遍历数据项中的数据是否为高维混合属性数据的频繁项,若是,则保留该遍历结果;若不包含,则舍弃该遍历结果,重新进行遍历^[15]。

(6)针对保留下来的数据频繁项集,利用式(19)和式(20)计算数据的支持度与置信度,将小于

最低支持度与置信度的频繁项作为数据的关联规则。

(7)关联规则的评估被解释为知识并被存储。

以上为基于 FP-growth 算法的高维混合属性数据挖掘的全部过程,满足最小支持和最小置信阈值的数据输出视为有效的关联规则。至此,完成对高维混合属性数据的挖掘。

2 实验论证分析

2.1 实验准备

对于提出的基于 FP-growth 算法的高维混合属性数据挖掘方法,结合仿真实验来说明算法执行的有效性。实验开展环境为 IOOM 局域网,采用 3 台独立计算机作为网络的分布式节点,每台电脑都配置了 CPU P4 2.0G、硬盘 3 GB 以及专业操作系统。网络主节点利用 PC 工作站替代,硬件和软件配置与上述相同。实验数据来自一个大型网络平台数据库,其中存储的数据量约有 36 万条,实验数据集构成如表 1 所示。

表 1 实验数据集构成

数据集	个数	条件属性个数	决策值类别数
zoo	101	17	7
Iris	150	4	3
Wine	178	13	8
Machine	209	7	7
Glass	214	9	2
Voting	435	16	3

基于以上数据集,将数据映射到高维空间的三个维度中,数据的传输延迟为 3 s,特征采样频率为 2.5 s,非高斯噪声的输出范围为-2~5Db。根据实验条件,对上述数据集的统计特征进行提取,得到特征量分布结果如图 1 所示。

以模拟实验中获得的数据统计特征量分布为测试对象,对数据的每个属性进行泛化处理,共得到 3~12 个属性值,每个属性值的最低支持度为 6%,最小置信度为 75%。

2.2 实验说明

对数据流时域波形进行扫描,计算 FP-树每个分支节点的支持度与置信度,并进行降序排列,剔除小于 6%支持度和 75%置信度的频繁项,并存储在频繁项表中。同时,舍弃频繁项前一节点和后一节点的项集,由此生成 FP-树与项表头。其中,

子节点为数据特征量的分布幅值,则构建的 FP-树如图 2 所示。

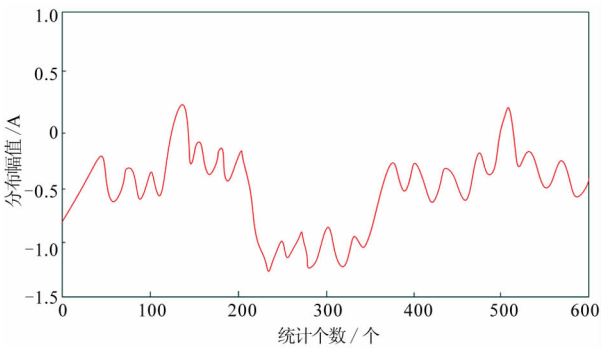


图 1 高维混合属性数据统计特征量分布

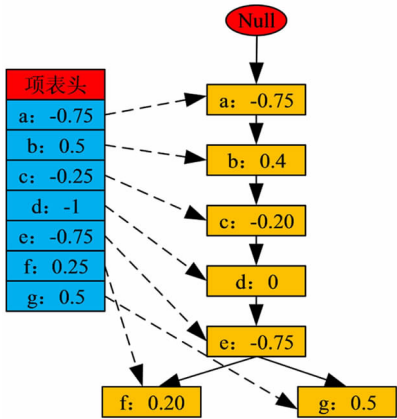


图 2 数据统计特征量生成的 FP 树

创建 FP 树根节点,将其标记为 T,并标记为 null。对数据流中的每个会话执行频繁的项目选择,并按项目标题中的顺序排序,然后将关联规则评估转换为知识并存储,便于算法执行。

2.3 算法加速对比实验分析

为分析本文算法在挖掘效率方面的性能,选取加速比这一评价指标进行评估。算法的加速比计算公式如下:

$$\alpha = \frac{q_i}{q_t} \tag{21}$$

其中, q_i 表示算法顺序执行的时间; q_t 表示当有 t 个节点时,并行算法的执行时间。

并将文献[1]提出的基于差分进化算法的数据挖掘方法(方法 1)、文献[2]提出的基于物联网边缘计算的数据挖掘方法(方法 2)作为对比组。在相同数据节点条件下,加速比越大,表明挖掘效率越高。三种算法挖掘时的加速比结果对比如图 3 所示。

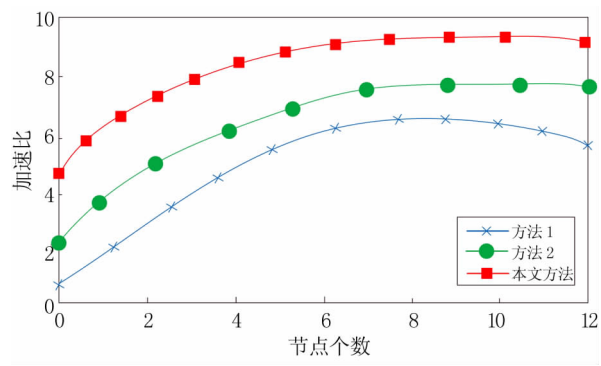


图3 不同数据挖掘算法的加速比性能对比

由图3可知,随着节点数量的增加,三种挖掘算法的加速比均呈先增加后下降的趋势。在节点数目为10时,本文方法的加速比达到最大值9.4,且挖掘加速比始终保持在4以上;而方法1与方法2的加速比最大值分别为5.8和7.5,本文方法与之相比,加速比分别提升了3.6和1.9。由此可见,利用本文方法对高维混合属性数据进行挖掘时,可有效提升挖掘速度,挖掘效率较高。

2.4 数据挖掘时间对比实验分析

为了进一步验证本文算法的挖掘效率,采用Webdocs.dat作为实验数据,该实验数据的大小为1448580 KB,包括1692082条事务记录,5267656个不同的项。在数据集中随机选取总数的10%、30%、50%、80%和90%,构建数据1、数据2、数据3、数据4和数据5。在集群结点固定,采取不同比例的该数据进行实验,在单机上运行方法1、方法2与本文方法,记录数据挖掘时间,对比挖掘效率。对比结果如图4所示。

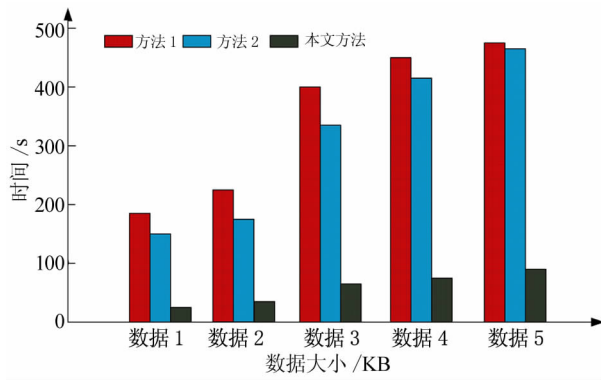


图4 不同算法的挖掘时间对比图

分析图4可知,当处理相同数量的数据时,单机运行本文方法对数据进行挖掘,挖掘时间总是远小于其他两种方法,主要是因为本文方法在对数据提取关联规则时能够完整保留数据结构的相似度,从而省去筛选无效规则的时间。实验结果表明,本

文提出的数据挖掘方法的挖掘效率更高,可应用于高维混合属性数据的挖掘工作中。

3 结 论

利用FP-growth算法设计了高维混合属性数据挖掘方法。通过固定算法对数据进行多角度去噪处理,基于FP-growth算法,实现高维混合属性数据挖掘。通过实验得出,利用本文数据挖掘技术,能够有效保障数据完整性,提升挖掘速度。在后续研究中,可继续优化所提方法,将其应用于各个领域,实现对多种类型数据的挖掘。

参考文献

[1] 蓓蓓. 基于差分进化算法的网络多属性大数据聚类挖掘方法[J]. 宁夏师范学院学报, 2021, 42(1): 91—97.

[2] 郑琳. 基于物联网边缘计算的数据挖掘方法研究[J]. 无线互联科技, 2022, 19(15): 140—142.

[3] 王伟. 基于机器学习的数据库异常信息挖掘方法[J]. 电子技术, 2022, 51(10): 24—25.

[4] 段晓萌, 王爽, 赵婷, 等. 基于FP-growth算法的用电异常数据挖掘方法[J]. 电子技术应用, 2020, 46(10): 47—50.

[5] 耿瑞焕. 面向船舶避碰的异常轨迹点数据动态挖掘方法[J]. 舰船科学技术, 2022, 44(22): 136—139.

[6] 李雪梅, 马文辉, 张春庆, 等. 多属性决策模型在网络大数据挖掘中的应用[J]. 中国新技术新产品, 2022(19): 55—57.

[7] 赵海燕, 杜丽娟, 刘琨, 等. 分布式光纤预警系统同质序列数据异常模式挖掘方法[J]. 激光杂志, 2022, 43(9): 134—138.

[8] 高天宇, 王庆荣, 杨磊. 粗糙集属性依赖度强化的应急数据挖掘模型[J]. 计算机工程与应用, 2021, 57(3): 87—93.

[9] 曹丽娜, 王霞, 周瑛. 基于模式匹配算法的空间属性数据挖掘仿真[J]. 计算机仿真, 2022, 39(9): 273—276.

[10] 李娟. 基于Hadoop云平台的空间属性数据挖掘技术研究[J]. 南京理工大学学报, 2022, 46(4): 419—426.

[11] 张庆昌. 基于人工智能的计算机网络异常数据挖掘方法[J]. 信息与电脑(理论版), 2022, 34(16): 31—33+37.

[12] 宫帅, 宋善坤. 多维关联规则的分布式能源系统数据挖掘方法研究[J]. 能源与环保, 2022, 44(10): 278—283.

[13] 唐建海. 基于Hadoop平台的网络安全趋势大数据挖掘算法[J]. 工业加热, 2022, 51(7): 67—70.

[14] 王营, 高琦, 李婷玉, 等. 基于改进FP-growth算法的售后服务数据挖掘[J]. 现代制造工程, 2021(6): 31—37.

[15] 程雅琼. 基于改进聚类算法的网络平台异常数据挖掘方法[J]. 长江信息通信, 2022, 35(4): 38—40.