

基于模糊关联规则的海量数据挖掘方法研究

毛晓菊

(商丘学院 计算机工程学院,河南 商丘 476000)

摘 要: 针对传统的数据挖掘方法一直存在计算时间复杂度较高,耗时长的问题,提出基于模糊关联规则的海量数据挖掘方法,首先通过对模糊理论、关联规则挖掘、模糊关联规则挖掘进行分析,实现对模糊关联规则的论述,在此基础上,将海量数据进行模糊化处理,建立模糊数据库,并对模糊数据库中模糊数据进行分类聚集,以及连续属性离散化处理,确定模糊数据库的频繁关联规则,从而实现基于模糊关联规则的海量数据挖掘方法.实验结果表明,所提方法能够降低数据挖掘消耗的时间及内存,且进行数据的挖掘支持度较高,为数据挖掘提出新的思路.

关键词: 模糊关联规则;海量数;数据挖掘

中图分类号: TP311

文献标识码: A

文章编号: 1000—7180(2018)02—0089—05

Research on Massive Data Mining Method Based on Fuzzy Association Rules

MAO Xiao-ju

(School of Computer Engineering, Shangqiu University, Shangqiu 476000, China)

Abstract: According to the traditional data mining methods have been calculated with high time complexity, time-consuming problem, proposed massive data mining method based on fuzzy association rules, based on fuzzy theory, association rules mining, mining fuzzy association rules analysis, implementation of fuzzy association rules is discussed, on the basis of this, the mass data are fuzzy with the establishment of fuzzy database, and to classify the aggregation of fuzzy data in fuzzy database, and discretization of continuous attributes, determine the frequent association rules in fuzzy database, so as to realize the data mining method based on Fuzzy Association rules. Experimental results show that the method proposed in this paper can reduce the time and memory consumed by data mining, and this method has more support for data mining, and puts forward new ideas for data mining.

Key words: fuzzy association rules; massive quantity; data mining

1 引言

随着计算机网络时代的到来,网络数据量越来越多,为了提高海量数据的利用效率,需要对海量数据进行挖掘^[1].通过数据挖掘从海量的、存在噪声的、模糊的数据中提取隐含在其中不为人知的但存在利用价值的有用信息^[2],通过数据挖掘,能够有效对数据中信息进行提取,提高数据的利用效率^[3].数据挖掘具有准确性和实时性,目前大多数海量数据挖掘方法通过从海量数据中发现隐含的规律性内容,提取有用信息,去除数据中的噪声及无用信息,

提高数据的质量,通过描述当前海量数据挖掘的方法,分析方法存在的问题,并指出本文所提的方法的优越性,并给出本文所提基于 Web 的数据挖掘方法实现过程,从而完成海量数据挖掘^[4].这种方法能够有效提高数据挖掘的可靠性,成为当前数据挖掘的主要方法^[5],引起了相关专家学者的广泛应用,随着大数据时代的到来,该课题成为研究的重点课题^[6].随着研究内容的深入,产生了许多与该课题相关的研究成果^[7].

文献[8]提出了一种基于关联规则映射的海量数据挖掘方法,首先对当前数据挖掘存在的问题

进行分析,阐述文章的必要性,然后对数据之间存在的关联映射关系进行分析,从而确定进行数据挖掘的关联规则,以相对误差方法为依据,利用关联因子提高数据挖掘的精度;以多维子空间中数据自检的关联性为依据,对子空间以及子空间中数据进行区分,从而实现对不同数据的挖掘.但这种方法存在挖掘过程复杂,不利于推广使用的问题.文献[9]提出基于相关子空间的海量数据挖掘方法,首先以 K 最近邻方法为依据,对数据进行聚集,利用数据属性值的稀疏因子生成数据矩阵,并计算数据因子对应的数据矩阵,以高斯误差函数为基础,对数据之间的关联性进行分析,从而实现数据挖掘.但这种方法存在数据挖掘的精度较低,可靠性较弱的问题.文献[10]提出了一种基于人工蜂群的海量数据挖掘方法,首先以小阿伯变换理论为依据,对数据进行变换,使数据标准化处理,降低数据挖掘的难度,然后利用关联系数对数据进行分类,并实现数据挖掘目标函数的构建,并利用人工蜂群算法对建立的目标函数进行优化,从而完成数据挖掘.但这种方法存在挖掘过程复杂,不利于操作的问题.

针对上述问题,本文提出了一种基于模糊关联规则的海量数据挖掘方法,首先对模糊关联规则进行分析,在此基础上,将海量数据进行模糊化处理,并对处理后的模糊数据进行分类聚集,以及连续属性离散化,实现基于模糊关联规则的海量数据挖掘方法.实验结果表明,本文所提方法进行数据挖掘消耗的时间较短,占用内存较少,且本文多提方法进行数据挖掘支持度较高,为数据挖掘的应用研究创造了有利条件.

2 基于模糊关联规则的海量数据挖掘方法研究

2.1 模糊关联规则

模糊理论是在普通集合理论的基础上,将取值范围从 $\{0,1\}$ 扩展到 $[0,1]$ 的概念理论,设定论域中存在元素 w 和集合 X ,则在普通集合理论中, w 与 X 的关系为

$$L(X, w, W) = \begin{cases} 1, w \in X \\ 0, w \notin X \end{cases} \quad (1)$$

在模糊集合理论中,设定 $f_x(w)$ 表示元素 w 对集合 X 的隶属度,其取值范围为 $[0,1]$,则 w 与 X 的关系为

$$L(X, w, W) = f_x(w), f_x(w) \in [0,1] \quad (2)$$

模糊集合通过将数据集合 $I = \{I_1, I_2, \dots, I_m\}$ 划

分成 K 类,并在每一类中设置一个聚类中心,设定 U 表示模糊隶属度矩阵,则模糊理论的目标函数可以表示为

$$J_S(U, k_1, \dots, k_K) = \sum_{i=1}^K \sum_{j=1}^m (u_{ij})^W d_{ij}^2 \quad (3)$$

式中, d_{ij} 表示第 i 个聚类区间的聚类中心与第 j 个数据之间的聚类, W 表示海量数据增量参数, u_{ij} 表示第 j 个数据隶属于第 i 个聚类中心隶属度,需要满足:

$$\sum_{i=1}^K u_{ij} = 1, \text{其中 } u_{ij} \in [0,1], \forall j = 1, 2, \dots, m \quad (4)$$

根据建立的目标函数,要保证目标函数取得最小值,从而完成模糊集合.

关联规则挖掘主要是由数据特征的前件 X 和后件 Y 组成的形如 $X \rightarrow Y$ 的规则,其具有支持度以及置信度两个参数,表示若 X 发生, Y 将以置信度的概率发生.关联规则显示的是数据与数据之间隐含的关系.关联规则挖掘就是发现数据之间这种隐含的联系.关联规则挖掘可以分为生成频繁项集和生成规则两大部分,首先选择待挖掘的海量数据库,通过参数设置实现生成频繁项集,在此基础上,生成关联规则,并将关联规则存储到关联规则库在,从而完成关联规则挖掘.

为了提高数据挖掘效果,在一般关联规则的基础上,通过将海量数据库 D 中数量型数据分段转换为布尔型数据,从而得到布尔型数据库,并将模糊集理论引入到数据库中,将布尔型数据库转换为模糊数据库 D_f ,在通过 *Apriori* 算法对模糊数据库中关联规则进行查找,完成模糊管理规则,因此模糊关联规则可表示成 $X \Rightarrow Y$ 蕴涵式,设定 i_j 为海量数据库集合中的元素,取值范围为 Δj ,且 Δj 与海量数据模糊集相关联,当 i_j 扩展到 Δj 的模糊集上,则用项目集 I 可得扩展项目集 I_f .

对于任意项目集 $X = \{X_1, X_2, \dots, X_p\} \subset I_f$,对于模糊数据库中第 i 条数据, X 的支持度可以表示为

$$D_{\text{sup}p_i}(X) = T(X_{1_i}, X_{2_i}, \dots, X_{p_i}) \quad (5)$$

式中, X_{j_i} 表示 X_j 在第 i 条数据中的模糊值; T 表示含量数据的广义三角模.

则关联规则的支持度可以表示为

$$D_{\text{sup}p_i}(X \Rightarrow Y) = \frac{\sum_{i=1}^n D_{\text{sup}p_i}(X \cup Y)}{|D_f|} \quad (6)$$

关联规则在整个模糊数据库中的蕴涵度可表示为

$$D_{imp}(X \Rightarrow Y) = \frac{\sum_{i=1}^n D_{imp_i}(X \Rightarrow Y)}{|D_f|} \quad (7)$$

通过上式论述,对模糊理论和关联规则挖掘进行分析,确定模糊关联规则挖掘的实现过程,完成对模糊关联规则的分析,为后续海量数据挖掘实现创造条件。

2.2 基于模糊关联规则的海量数据挖掘方法

基于模糊关联规则的海量数据挖掘就是将模糊集理论引入关联规则中,形成模糊关联规则,并通过模糊关联规则实现海量数据挖掘。为实现基于模糊关联规则的海量数据挖掘,首先需要对具有最小模糊支持度的频繁项集进行查找,并利用查找到的频繁项集生成用于海量数据挖掘的模糊关联规则,从而完成基于模糊关联规则的海量数据挖掘。首先,选取海量数据挖掘的模糊关联规则,其形如 $X \Rightarrow y$, 其中, X 表示海量数据的属性值集合, y 表示海量数据的类标号;利用启发式策略或加权分类规则实现海量数据分类器构建。

设定海量数据分类规则的后件仅有一个类, A_q 表示海量数据属性 x_i 对应的一个模糊项, CF_{qm} 表示模糊关联规则 R_q 将海量数据样本预测为类 C_{qm} 的置信度。

本文中,分类器的构建是通过分析每个海量数据分类规则的权重实现的,每个规则的权重可以表示为

$$\omega_q = c(A_q \Rightarrow C_q) - c_{2nl} \quad (8)$$

式中, c_{2nl} 满足公式:

$$c_{2nl} = \max\{c(a_q \Rightarrow h) \mid h = 1, 2, \dots, M; h \neq C_q\} \quad (9)$$

设定 C_m 表示通过基分类器 S 预测海量数据样本 X_p 的类标号,则有:

$$\mu_{A_q}(X_p) \cdot \omega_m = \max\{\mu_{A_q}(X_q) \cdot \omega_q \mid R_q \in S\} \quad (10)$$

为了对海量数据中每个样本进行加权,首先通过当前的分类器对海量数据样本进行预测,确定当前分类器分类的准确性,并给予错分样本更高的权重,设定 ω_i 表示海量数据样本 X_i 的权重,并存在 $\sum_{i=1}^m \omega_i = 1$, 则得到样本的支持率和置信率可以表示为

$$s(A_q \Rightarrow C_q) = \sum_{X_p \in C_q} \mu_{A_q}(X_p) \omega_p \quad (11)$$

$$c(A_q \Rightarrow C_q) = \frac{\sum_{X_p \in C_q} \mu_{A_q}(X_p) \omega_p}{\sum_{p=1}^m \mu_{A_q}(X_p) \omega_p} \quad (12)$$

基于模糊关联规则的海量数据挖掘方法是通过最小支持度方法,通过给予海量数据中少数类相对低的最小支持度进行数据挖掘,有效减低由于数据不平衡对分类器预测精度差问题,提高分类的准确性,从而提高海量数据挖掘效果。设定 $\min \sup_i$ 表示类 c_i 的分类规则最小支持度, base_minsup 表示支持度基数, $\text{freqDistr}(c_i)$ 表示类 c_i 占海量数据的比例,则有:

$$\text{base_min sup}_i = \text{base_minsup} * \text{freqDistr}(c_i) \quad (13)$$

通过上述论述,完成海量数据分类器构建,实现海量数据分类。在此基础上,对海量数据进行聚类,考虑到应用在数据挖掘过程中,用基于模型的聚类方法。

对海量数据进行整理,从而保证海量数据挖掘的准确性。由于海量数据具有连续性,因此利用模糊离散化进行数据处理,模糊离散化是模糊数据挖掘中的重要内容,其可以将应用于离散空间的方法扩展到应用于连续空间,从而实现现实问题的处理,并且给予连续属性的离散化以降低数据挖掘的难度,及时间上的消耗。因此在海量数据库中,通过对海量数据属性进行离散化处理,来提高数据样本提取的准确性,降低基于模糊关联规则的海量数据挖掘的难度,提高海量数据挖掘效率。

连续属性模糊离散化就是将特定的连续属性值域的数据划分为若干个模糊离散化的划分点,得到若干个属性区间,并通过隶属度函数对每个属性区间进行赋值,从而实现对属性区间的描述。基于模糊关联规则的海量数据挖掘方法过程如图1所示。

3 实验结果与分析

为了证明本文提出的基于模糊关联规则的海量数据挖掘方法的可行性,以 Intel P4 2 GHz 处理器为硬件环境, MATLAB2008a 为实验平台,通过对 web 网络数据库中数据进行挖掘,分析其挖掘性能,从而完成实验。

首先针对不同数据集,将使用本文方法和文献[7]、文献[9]方法进行数据挖掘所消耗的时间(min)进行对比,对比结果见图1。

通过图2可以看出,本文所提方法进行数据挖掘消耗时间最少,且本文所提方法进行数据挖掘消

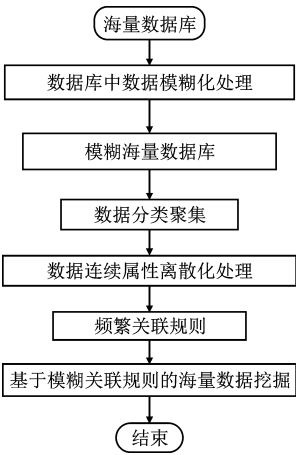
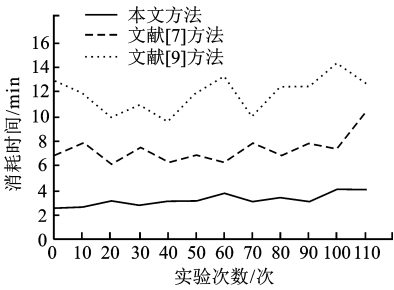


图 1 基于模糊关联规则的海量数据挖掘流程

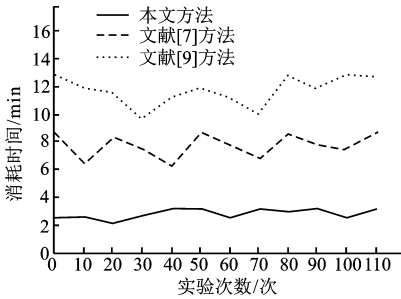
耗时间折线较文献[7]、文献[9]消耗时间折线平缓, 由于本文是对同一数据集进行多次实验, 因此消耗时间折线波动越平缓, 对数据集进行数据挖掘的时间波动越小, 方法越为稳定, 因此本文所提方法较为稳定, 由于本文所提方法在进行海量数据挖掘的过程中, 对数据进行了分类聚集, 降低了数据挖掘的时间消耗, 时间消耗较为平缓。

分别采用本文方法和文献[8]、文献[9]方法进行数据挖掘, 对比进行数据挖掘需要消耗的计算机内存, 通过实验, 得到三种方法进行数据挖掘需要消耗的计算机内存空间(MB)对比, 对比结果如图 3 所示。

通过图 3 可以看出, 本文所提方法进行数据挖



(c)数据集 Connect



(d)数据 Pumsb

图 2 不同数据集进行数据挖掘消耗时间对比

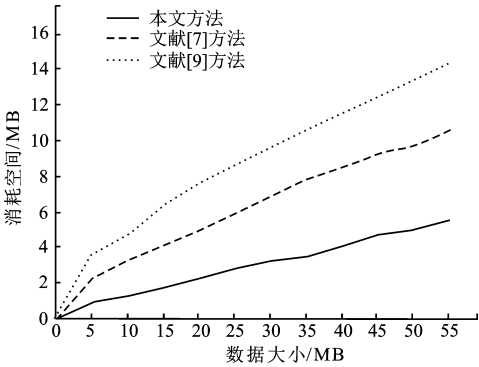
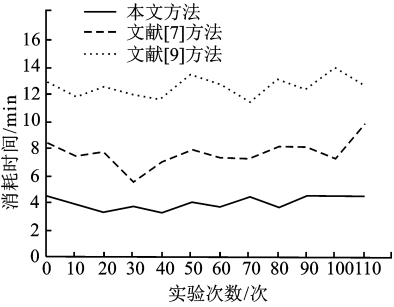
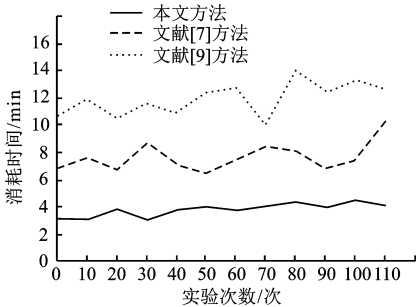


图 3 三种方法进行数据挖掘消耗空间对比



(a)数据集 T40I10D10K



(b)数据集 Kosarak

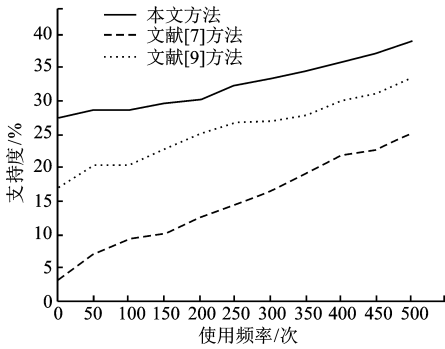


图 4 三种方法进行数据挖掘支持度对比

掘占用的空间小于文献[8]和文献[9]方法进行数据挖掘所占用的空间, 由于本文所提方法进行数据挖掘, 充分考虑了模糊关联规则优越性, 因挖掘所占用内存空间较小。

最后对比三种算法的支持度(%), 对比结果如

图 4 所示。

从上图可以看出,本文所提方法的支持度最高,由于本文进行数据挖掘,对数据的属性进行了充分分析,因此数据挖掘的效果较好,支持度较高。

综上所述,本文所提方法进行海量数据挖掘,挖掘消耗的时间较短,占用内存较少,且数据挖掘的支持度较高,因此本文所提方法进行数据挖掘效果较好。

4 结束语

随着大数据时代的到来,数据信息量越来越多,对海量数据进行挖掘,分析其中隐含的信息,可以提高数据的利用价值,从而提高数据利用率,针对传统方法存在的数据挖掘效率低,成本高问题,本文提出了一种基于模糊关联规则的海量数据挖掘方法,实验结果表明,使用本文所提方法进行数据挖掘,挖掘速度较快,占用内存空间较低,为该领域的研究发展创造条件。

参考文献:

[1] 李骏骁. 多层差异网络深度入侵数据挖掘方法研究[J]. 计算机仿真, 2015, 32(4): 235-238.

[2] 汤恒耀, 胡志华. 基于链距离估计的非显著特征数据挖掘算法[J]. 科技通报, 2015, 31(6): 142-144.

[3] 任高举, 白亚男. 多媒体智能教学系统中特定数据挖掘

方法研究[J]. 电子设计工程, 2016, 24(11): 4-7.

[4] 张继荣, 王向阳. 基于 XML 数据挖掘的 Apriori 算法的研究与改进[J]. 计算机测量与控制, 2016, 24(6): 178-180.

[5] 梅毅, 熊婷, 罗少彬, 等. 复杂属性环境下 NoSQL 分布式大数据挖掘方法研究[J]. 科学技术与工程, 2017, 17(9): 239-243.

[6] 陈炎龙, 段红玉. 基于改进 Hadoop 云平台的海量文本数据挖掘[J]. 湖南师范大学自然科学学报, 2016, 39(3): 84-88.

[7] 贾克斌, 李含婧, 袁野. 基于 Apriori 算法的数据挖掘在移动医疗系统中的应用[J]. 北京工业大学学报, 2017, 43(3): 394-401.

[8] 唐晓东. 基于关联规则映射的生物信息网络多维数据挖掘算法[J]. 计算机应用研究, 2015, 32(6): 1614-1616.

[9] 李永红, 张继福, 荀亚玲. 相关子空间中的局部离群数据挖掘算法研究[J]. 小型微型计算机系统, 2015, 36(3): 460-465.

[10] 宋正江, 李晓晨, SONGZheng-Jiang, 等. 基于人工蜂群的 WSN 故障数据挖掘算法[J]. 四川大学学报(自然科学版), 2015, 52(1): 51-56.

作者简介:

毛晓菊 女, (1982-), 硕士研究生, 讲师. 研究方向为数据挖掘. E-mail: meizhuan1001@163. com.

(上接第 88 页)

[6] 潘兴境. 集成电路平均传输延迟时间的两种测试方法[J]. 化工自动化及仪表, 1979, 14(4): 77-83.

[7] 朱明程, 熊元姣. ACTEL 数字系统现场集成技术[M]. 北京: 清华大学出版社, 2004.

作者简介:

杜涛 男, (1976-), 博士, 讲师. 研究方向为大规模集成电路设计与验证. E-mail: dutao@uestc. edu. cn.

李威 男, (1959-), 教授, 博士生导师. 研究方向为可编程

逻辑器件、集成电路可靠性技术.

晁醒 男, (1993-), 硕士研究生. 研究方向为大规模集成电路设计.

吴方明 男, (1991-), 硕士研究生. 研究方向为大规模集成电路设计.

吴华刚 男, (1988-), 硕士研究生. 研究方向为大规模集成电路设计.

刘丹 女, (1990-), 硕士研究生. 研究方向为大规模集成电路设计.