文章编号：2095-2163(2024)03-0207-05 中图分类号：TP391 文献标志码：A

混合属性网络多维多层关联数据智能挖掘算法

段雪莹
(吉林警察学院 信息工程系，长春 130117)

摘 要：针对传统关联数据挖掘算法，强项集挖掘后产生大量候选项集，导致挖掘耗时长、挖掘精度低等问题，提出一种混合属性网络多维多层关联数据智能挖掘算法(Multidimensional Multilayer Associative Data Intelligent Mining Algorithm，MMAD-IM)。计算混合属性网络中随机数据到簇中心的距离，将目标数据分配到距离簇中心最近的簇中，使簇中心固定，完成混合属性网络数据的聚类分析。从聚类完成的数据中提取出有效的基本频繁向量，同时计算数据的候选项集，对哈希表进行扫描，利用改进 Apriori 算法完成强项集挖掘。以此为基础构建空间关系，获取近似区域与近似点之间的距离，形成待挖掘数据并计算数据的隶属度数值，完成智能挖掘。实验结果表明，所提算法具有较好的数据聚类效果，强项集挖掘后剩余的候选项集数量较少，整体数据挖掘耗时远低于传统算法，挖掘精度高达 90%。

关键词：多维多层关联数据；聚类；基本频繁向量；强项集；挖掘

Intelligent mining algorithm for multi-dimensional and multi-layer association data in hybrid attribute networks

DUAN Xueying
(Department of Information Engineering, Jilin Police College, Changchun 130117, China)

Abstract: For traditional association data mining algorithms, strong itemset mining generates a large number of candidate itemsets, leading to long mining time, low mining accuracy, and other problems. A hybrid attribute network multi-dimensional multi-layer association data intelligent mining algorithm is proposed. The distance from the random data in the mixed attribute network to the cluster center is calculated, and the target data is assigned to the cluster closest to the cluster center, so that the cluster center is fixed and the clustering analysis of the mixed attribute network data is completed. The effective basic frequent vectors are extracted from the data completed by clustering, and the candidate itemsets of the data are also calculated, the hash table is scanned, and the strong itemset mining is completed by using the improved Apriori algorithm. Based on this, the spatial relationship is constructed, the distance between the approximate region and the approximate point is obtained, the data set to be mined is formed and the affiliation value of the data is calculated to complete the intelligent mining. The experimental results show that the proposed algorithm has better data clustering effect, the number of remaining candidate itemsets after strong itemset mining is smaller, the overall data mining time is much lower than the traditional method, and the mining accuracy is up to 90%.

Key words: multi-dimensional multi-layer association data; clustering; basic frequent vector; strong itemsets; mining

0 引 言

人类生产和生活中需要运用大量的数据，且呈现大规模持续增长趋势，这些数据类型多样、数量庞大，构成了多维度、多层次的数据混合属性体系[1-4]。数据挖掘是人类一种最基本的学习和知识发现模式，混合属性体系是数据挖掘的基础[5-8]，对混合属性网络数据进行挖掘处理，能够把低水平的数据映射为更有用、更紧凑的数据类型，方便网络数据的应用[9-12]。

由于混合属性网络数据具有海量性、高维性等特点，使得数据挖掘更具挑战性，因而引起众多学者的广泛关注。文献[13]从不确定的数据中对频繁项集进行挖掘，同时保证挖掘结果能够符合差分隐私的需求，从而获取用户的敏感信息，并通过理论和实验分析内容，对该方法进行了验证，但该方法存在挖掘耗时较长的问题。文献[14]对隐私保护数据挖掘技术进行了分析，基于粒度计算的方式对高效隐私保护频繁模式挖掘算法(Efficient mining associations with secrecy konstraints，EMASK)进行优

基金项目：吉林警察学院院级科研项目(jykyzd202404)。

作者简介：段雪莹(1979-)，女，副教授，主要研究方向：计算机应用技术、大数据、人工智能。Email：duanxueyingxy2020@126.com

收稿日期：2023-04-24

化,解决增量式数据库中频繁项集的计算问题,在完成隐私保护的条件下,展开了数据挖掘研究,但在数据挖掘精度方面还有待提高。文献[15]提出一种高效的多类型数据挖掘算法,根据算法所生成的规则构建了数据分类器,给出了多类数据的直观因果关系图,并通过实验验证了所提算法挖掘信息的有用性,但挖掘精度欠佳。

针对上述研究存在的问题,本文提出一种混合属性网络,多维多层关联数据智能挖掘算法(Multidimensional Multilayer Associative Data Intelligent Mining Algorithm, MMADIM),对混合属性网络数据进行聚类分析,以聚类后的数据为对象,从多维频繁基本向量和强项集方面加以分析,进行数据挖掘研究,以期寻求更高效的数据挖掘算法。

1 混合属性网络数据聚类

聚类分析是数据挖掘中的重要技术之一[16-17],是一种数据划分方式。利用数据聚类分析,能够获取数据潜在结构,是完成数据挖掘的先决步骤之一,本节将对混合属性网络数据进行聚类分析。

将需处理的目标混合属性网络数据表示为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$,$n$ 为数据数量;每个数据对象包含 i 个属性,用 $m_1,m_2,\cdots,m_i$ 表示;每个数据对应的属性取值范围,可用值域 $Dom(Z)$ 进行表示。本文研究的混合属性网络数据可划分为 2 个属性值域:分类属性值域和数值属性值域[18-19]。其中,数值属性值域通常情况下为连续数值形式,分类属性值域可表示为:

$$Dom(Z_i)=\{z_i^1,z_i^2,\cdots,z_i^t\} \tag{1}$$

其中,t 表示在混合属性网络数据 X 中,属性 i 对应的属性值数量。

通常目标数据 X 可以表示为属性值对的并集形式,表示为:

$$X=[m_1=z_i^1]\wedge[m_2=z_i^2]\cdots\wedge[m_i=z_i^t] \tag{2}$$

运用目标函数将数据 X 划分到 l 个簇中,将目标函数固定为特定函数,表示为:

$$H(x_n)=\sum_{n=1}^{l}p(x_n)d(x_n) \tag{3}$$

其中,$d(x_n)$ 表示数据对象到簇 l 中心的距离,$p(x_n)$ 表示数据目标划分函数。这里,$d(x_n)$ 可由式(4)计算求出:

$$d(x_n)=X\sum_{n=1}^{l}(s_n-q)^2 \tag{4}$$

其中,q 表示簇中心,s_n 表示随机数据对应的分类属性值。

根据式(4)计算混合属性网络数据 X 中随机数据到簇中心的距离,将目标数据分配到距离簇中心距离最近的簇中,分配结束将再次更新簇中心,以此循环,直至计算得出的数据,无法对簇中心产生影响、即簇中心固定,则表示完成了混合属性网络数据的聚类分析。

2 多维多层关联数据智能挖掘算法

2.1 多维频繁基本向量提取

数据挖掘过程中,所有的非频繁向量都不会出现在多维多层数据序列中,因此多维多层关联数据智能挖掘的基本步骤是提取频繁基本向量,从数据库中寻找出有效的基本频繁向量。

多维关联规则[20]是指具有多个属性的规则,不仅仅针对单个维度的频繁项集进行挖掘,所以在挖掘过程中,需要保证基本向量的底部元素为多个叶子节点的集合,底部向量集合的表达形式为 $\boldsymbol{A}=\{A_{1i},A_{2i},\cdots,A_{ki}\}$,这里,$k$ 为叶子节点数量,所有由底部向量组合而成的向量集表示为 J。随机给定 2 个隶属于向量集 J 的基本向量:

$$q=\boldsymbol{A}[q_{1i},q_{2i},\cdots,q_{ki}] \tag{5}$$

$$q'=\boldsymbol{A}[q'_{1i},q'_{2i},\cdots,q'_{ki}] \tag{6}$$

将 q 和 q' 进行并连接,记作 $q\vee q'$,表示为:

$$q\vee q'=(q_{1i}\vee q'_{1i},q_{2i}\vee q'_{2i},\cdots,q_{ki}\vee q'_{ki}) \tag{7}$$

依据式(7)的方式递推,可生成隶属于向量集 J 的所有基本向量,计算得出其中每一个基本向量的支持度,则可生成多维频繁基本向量。

2.2 强项集挖掘

在上述多维频繁基本向量提取的基础上,对强项集挖掘进行分析。在传统的单一数据挖掘算法基础上进行改进,采用扩展 Apriorii(Extend_Apriori)与基于扩展哈希的 Apriori(Extendhash_Apriori)结合的 Apriori 改进算法,对强项集进行挖掘。在研究中,Extend_Apriori算法能够依据前几项待挖掘数据关联规则,获取含有原查询项的关联规则,构造规则库,确定最低支持度阈值,实现查询扩展。采用 Extendhash_Apriori 算法挖掘中,首先压缩数据,并对数据进行整数化处理,然后对每个数据标定固定编码,通过建立数据映射关系制成哈希表,在对强项集挖掘中,只需扫描哈希表,无需对整个数据库进行扫描。

综合上述 2 种算法的优势,本文提出一种改进 Apriori 算法,利用多维多层关联数据智能挖掘算法

挖掘强项集时，利用 Extend_Apriori 算法实现查询扩展，过程中候选项集前两项的数据扫描过程最为重要，能决定数据挖掘的最终效果。

候选项集越少，挖掘效果越好。为获取一个较少的候选项集，改进 Apriori 算法首先对候选项集进行过滤处理，在扫描计数过程中，预先收集候选项集信息，对应哈希桶对每一个候选项集进行计数。候选项集计数算法如下：

(1)在一维空间中对所有数据进行计数，赋予每个数据对应的地址，表示为：

$$H = \{\{f_k q_{ki}\} \mid (q_{ki} \in Maxspan)\} \tag{8}$$

其中，$Maxspan$ 表示一维空间，f_k 表示参考数据地址。

(2)根据强项集产生候选项集。

(3)假设已经获取强项集，则在候选项集产生后，需通过合并和过滤后返回强项集。合并过程可用式(9)进行表示：

$$B_{ki} = H\{f_1(q_{1i}), f_2(q_{2i}), \cdots, f_k(q_{ki})\} \tag{9}$$

其中，B_{ki} 表示拓广的强 k 项集，$f_1(q_{1i})$、$f_2(q_{2i})$、$f_k(q_{ki})$ 分别表示不同参考地址的候选项集。

利用 τ 表示关联系数，$f_k(q_{ki})$ 的计算公式为：

$$f_k(q_{ki}) = \begin{cases} 0, & f_k = 0 \\ B_{ki}\tau, & f_k \neq 0 \end{cases} \tag{10}$$

利用式(10)可知，拓广候选项集中参考数据地址会存在非零情况，影响候选项集取值，因此在强项集挖掘中需要对包含非零参考数据地址的候选项集进行处理，确保全部候选项集的参考数据地址为 0。

基于上述候选项集计数算法，对利用 Extend_Apriori 算法构建的哈希表进行扫描。首先对前 $k-1$ 次扫描获取的拓广强 $k-1$ 项集 Q_k 进行处理，生成 k 项的候选项集 H_k；再对哈希表进行重复扫描，对每一个候选项集进行计算，获取拓广的候选项集集合，生成拓广的强 k 项集。不断重复上述步骤，直至候选项集 H_k 为空，完成强项集挖掘。

3　数据挖掘实现

本节以多维频繁基本向量提取和强项集挖掘结果为基础，对混合属性网络多维多层关联数据智能挖掘进行深入分析。

按照多维频繁基本向量和强项集的分布以及重要程度，将混合属性网络数据进行排序，并构建一种空间关系，采用近似区域方法对空间关系进行表示。

假设近似区域为 g，其最大延伸范围为 j，则 g 的隶属函数可表示为：

$$g(j) = k\left(1 - \frac{d'(g_1 + g_2)}{j}\right) \tag{11}$$

其中，g_1 和 g_2 分别表示近似区域中的固定点；k 表示延伸系数；d' 表示欧式距离。

近似区域 G 由近似点的强项集集合组成，若 G 为固定区域，则 G 的隶属函数可称为 G 的核，可由式(12)表示为：

$$G(j) = k\left(1 - \frac{d'd''}{j}\right)g \tag{12}$$

其中，d'' 表示近似区域与近似点之间的距离，计算公式如下：

$$d'' = \begin{cases} 0, & j \in G \\ \min(jd'), & j \notin G \end{cases} \tag{13}$$

依据近似区域与近似点之间距离的计算结果，按照距离的优先排序，可获取混合属性网络数据之间的关联属性。随机选取一个多维多层关联数据属性值，赋予其固定概率为 β，若其他数据属性值隶属于概率 $(0, \beta)$ 之间，则认为该数据属性存在于待挖掘数据关联规则中；若其他数据属性值不隶属于概率 $(0, \beta)$，则认为该数据属性不存在于待挖掘数据关联规则中，由此形成待挖掘数据。

对待挖掘数据中各数据的隶属度进行计算，这一过程可被描述为式(14)~(16)：

$$u_1 = 1 - d''u_{a1} \tag{14}$$

$$u_2 = d''u_{a1}u_{a2}u_n \tag{15}$$

$$u_n = d''u_{a1}(1 - u_{a2})(1 - u_{an}) \tag{16}$$

其中，u_{a1}、u_{a2} 和 u_{an} 分别表示待挖掘数据中不同空间的数据，u_1、u_2 和 u_n 分别表示对应的隶属度数值。

根据隶属度数值计算结果，判断数据项之间的关联规则。若存在关联规则，则输出数据，并将其还原为原始属性值，重复迭代操作，直至待挖掘数据全部输出，则可完成混合属性网络多维多层关联数据智能挖掘。

4　实验结果与分析

为验证 MMADIM 算法的性能，选取 60 个混合属性样本进行算法性能测试分析，数据维数设置为 3~15，数据来源于 UCI 数据集。

4.1　数据聚类准确率分析

MMADIM 算法分析过程中，首先进行数据的聚类分析，为后续有效的数据挖掘提供基础保障。对数据聚类结果质量评估中，通常使用的一种评估标

准是聚类准确率,定义如下:

$$r = \frac{\sum_{a=1}^{k} s_a}{n} \tag{17}$$

其中,s_a 表示第 a 个数据簇中目标数据数量。

根据式(17)对 MMADIM 算法进行度量,聚类准确率越高,表示聚类效果越好,越有利于完成数据挖掘。实验分析过程中,为保证数据分析结果的准确性,使聚类过程运行 100 次,取数据聚类准确率的平均值。

对 60 个混合属性样本进行聚类测试结果见表 1。

表 1 聚类测试结果

Table 1 Cluster test results

混合属性样本	聚类准确率/%
10	89
20	86
30	91
40	85
50	92
60	90

由表 1 可以看出,在 60 个混合属性样本测试过程中,MMADIM 算法的数据聚类准确率均能够获取较高数值,最高准确率为 92%,表明 MMADIM 算法数据聚类处理的效果较好。

4.2 候选项集剩余数量对比

采用 Extend_Apriori 和 Extendhash_Apriori 两者结合的改进 Apriori 算法对强项集进行挖掘。在此过程中,候选项集是完成强项集挖掘的重要指标,候选项集越少,数据挖掘效果越好。为验证 MMADIM 算法强项集挖掘效果,将 MMADIM 算法与文献[13]及文献[14]算法的候选项集的剩余数量进行对比测试,结果如图 1 所示。

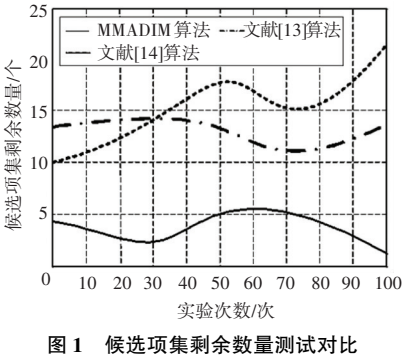

图 1 候选项集剩余数量测试对比

Fig. 1 Comparison of remaining quantity tests for candidate set

由图 1 可见,文献[13]和文献[14]采用的算法在处理同样的数据集时,产生的候选项集剩余数量较 MMADIM 算法高,表明这 2 种算法在候选项集的过滤和管理上不如 MMADIM 算法有效。文献[13]和文献[14]的算法在候选项集管理上有所不同,比如,没有使用哈希技术或者使用的过滤策略较为简单,可能是导致其在效率和效果上不如 MMADIM 算法的原因。MMADIM 算法在运行过程中产生的候选项集剩余数量始终保持在较低水平(不超过 5 个),这一结果强有力地证明了 MMADIM 算法在减少候选项集方面的有效性,从而提高了数据挖掘的效率。相比之下,文献[13]和文献[14]的算法在整个实验过程中产生的候选项集剩余数量波动较大,说明这些算法在处理复杂度和效率上可能存在缺陷。通过改进 Apriori 算法,并结合哈希技术,成功地降低了候选项集的数量,提高了数据挖掘的效率和效果。

4.3 整体数据挖掘耗时对比

根据多维频繁基本向量提取分析可知,多维多层数据序列中不会出现非频繁向量,因此选取其中的所有数据向量为研究对象,对 MMADIM 算法的多维频繁基本向量提取时间进行测试,测试结果见表 2。

表 2 多维频繁基本向量提取时间

Table 2 Multi-dimensional frequent basic vector extraction time

数据维数	提取时间/s
3	1.23
6	2.56
9	2.98
12	3.56
15	3.92

分析表 2 可知,在数据维数为 0~15 范围时,本文多维频繁基本向量提取时间控制在 3.92 以内,主要原因是在多维频繁基本向量提取过程中,对每一个生成向量的支持度进行计算,以支持度计算结果为依据完成向量提取,节约了提取耗时。

为进一步对表 2 结果进行验证,在不同数据维数条件下,将 MMADIM 算法与文献[13]和文献[14]算法的整体数据挖掘耗时进行对比,结果如图 2 所示。

根据图 2 整体数据挖掘耗时对比结果可知,随着数据维数的变化,整体数据挖掘耗时的变化较为明显。当数据维数较小时,基本频繁向量的生成数量较少,对应数据处理时间也较少,随着数据维数的增加,3 种算法的耗时均呈上升趋势。然而,在本实

验设定的最大数据维数 15 条件下，MMADIM 算法的挖掘耗时为 5.0 s 左右，文献[5]算法的挖掘耗时接近 10.0 s 左右，文献[6]算法的挖掘耗时接近 11.0 s左右，均高于 MMADIM 算法的 2.0 倍，表明 MMADIM 算法数据挖掘性能较好。实验结果充分说明了 MMADIM 算法在处理大规模和高维度数据集时的优越性。其耗时远少于文献[13]和文献[14]所提算法，这是由于 MMADIM 算法更加精确地定位和减少了频繁项集的搜索范围。

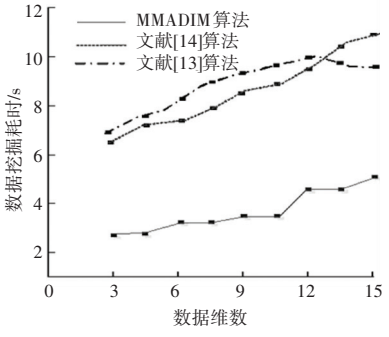

图 2　整体数据挖掘耗时对比

Fig. 2　Comparison of overall data mining time

4.4　数据挖掘精度分析

对 MMADIM 算法与文献[5]，文献[6]算法的数据挖掘精度进行对比分析，结果如图 3 所示。

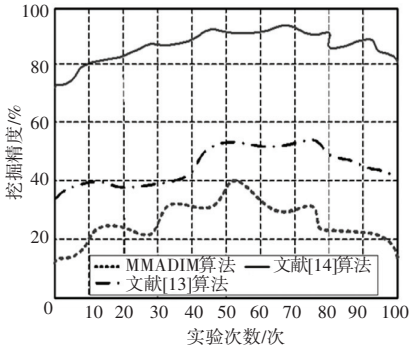

图 3　数据挖掘精度对比

Fig. 3　Comparison of data mining accuracy

根据图 3 可以看出，MMADIM 算法的挖掘精度可高达 90%，显著高于文献[13]和文献[14]算法，MMADIM 算法凭借其在数据挖掘精度上的优势，证实了改进后的 Apriori 算法在挖掘强项集方面的有效性。提高精度的同时，该算法仍保持了良好的计算效率，提高了整体数据挖掘精度。

5　结束语

本文提出一种混合属性网络多维多层关联数据智能挖掘算法。首先对混合属性网络数据进行聚类分析；其次，提取基本频繁向量，挖掘强项集；最后，通过对待挖掘数据中数据的隶属度计算，完成多维多层关联数据智能挖掘。实验结果表明，所提算法具有较好的数据挖掘性能，远远优于传统算法。但本研究还存在一定的不足，在以后的研究中，需要对数据的多维多层关联属性做进一步分析，并对数据挖掘展开深入研究，为数据处理技术的发展提供参考依据。

参考文献

[1] 陆江东，郑奋，戴卓臣. 异构大数据网络的多维关联细粒度数据挖掘算法[J]. 计算机系统应用，2018，27(3)：56-69.
[2] 温炜，刘媛媛，杨瑞. 基于数据挖掘算法的电力行业智能培训服务质量评估方法[J]. 湖北农业科学，2023，62(S1)：221-225.
[3] 杨晓燕. 基于关联规则挖掘的企业财务大数据智能整合方法[J]. 中国管理信息化，2023，26(21)：61-64.
[4] 阮诗迪，刘欣然，张雄宝，等. 智能电网海量信息处理中的关联规则与数据流挖掘研究[J]. 微型电脑应用，2023，39(7)：61-64.
[5] 章新友，徐华康，唐琍萍，等. 基于策略模式的中医药数据智能挖掘平台设计与应用[J]. 科学技术与工程，2023，23(14)：5946-5954.
[6] 宋煜，江志凌，刘艳超. 关联规则挖掘方法在输送线烟箱缺条智能检测中的应用[J]. 微型电脑应用，2023，39(4)：202-204，208.
[7] 欧阳烽. 基于数据挖掘的图书智能采购模式分析[J]. 电脑与信息技术，2023，31(1)：50-53.
[8] 张欣然，宋绍成，王娜. 突发事件监测防控中危机情报的智能挖掘研究[J]. 信息技术与信息化，2022(10)：196-199.
[9] 孙丰杰，王承民，谢宁. 面向智能电网大数据关联规则挖掘的频繁模式网络模型[J]. 电力自动化设备，2018，38(5)：110-116.
[10] 朱国进，凌晓晨. 基于关联规则挖掘的 OJ 推荐方法[J]. 智能计算机与应用，2018，8(2)：20-24.
[11] 郭鹏飞，李海霞，常海艳，等. 基于大数据的海上目标隐性关联规则挖掘方法[J]. 网络安全与数据治理，2023，42(S1)：71-77.
[12] 张庆，李梦，于晓涵. 基于 Apriori 算法的糖尿病患者用药规律关联规则挖掘分析[J]. 中国卫生统计，2023，40(5)：687-691.
[13] 丁哲，秦臻，秦志光. 基于差分隐私的不确定数据频繁项集挖掘算法[J]. 计算机应用研究，2018，35(7)：28-32.
[14] 程舒通，徐从富，但红卫. 增量式隐私保护数据挖掘研究[J]. 计算机应用研究，2018，35(7)：242-245.
[15] 张新英，付川南. 一种高效的多类型数据挖掘算法[J]. 中国电子科学研究院学报，2017，12(4)：359-364.
[16] 张翩，毛澍，李彦庆，等. 利用数据挖掘技术改进 TCP CUBIC 拥塞控制算法[J]. 计算机应用研究，2018，35(10)：170-173.
[17] WANG Tiexing，LI Qunwei，BUCCI D J，et al. K-medoids clustering of data sequences with composite distributions[J]. IEEE Transactions on Signal Processing，2019，67(8)：2093-2106.
[18] LU Canyi，FENG Jiashi，LIN Zhouqi，et al. Subspace clustering by block diagonal representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2019，41(2)：487-501.
[19] 丁继红，刘华中. 大数据环境下基于多维关联分析的学习资源精准推荐[J]. 电化教育研究，2018，6(2)：53-59.
[20] 邹晖，李金灿，卢万平. 基于多维关联规则的用电负荷智能预测方法[J]. 电子设计工程，2024，32(5)：122-126.