基于概念格的多值属性关联规则挖掘

郭晓波[1,2,3] 赵书良[1,2,3] 王长宾[1] 赵娇娇[1,2,3] 刘军丹[1,2,3]
(河北师范大学数学与信息科学学院 石家庄 050024)[1]
(河北省计算数学与应用重点实验室 石家庄 050024)[2]
(河北师范大学移动物联网研究院 石家庄 050024)[3]

摘 要 针对传统关联规则挖掘算法不利于用户选择关键数据进行分析、无法处理多值属性数据及效率低下等问题，提出了基于KAF因子和CHF因子的Apriori改进算法来进行多值属性关联规则挖掘，运用概念格理论对多值属性数据进行了重新定义和分类；建立了数据挖掘参数调整机制，以提高算法挖掘效率，方便用户选择关键属性值进行规则挖掘分析。结合某省全员人口数据对算法进行了具体实现和分析。实验结果表明，算法性能具有较大提高。

关键词 多值属性，概念格，关联规则，Apriori
中图法分类号 TP391.1 **文献标识码** A

Multi-valued Attribute Association Rules Mining Based on Concept Lattice

GUO Xiao-bo[1,2,3] ZHAO Shu-liang[1,2,3] WANG Chang-bin[1] ZHAO Jiao-jiao[1,2,3] LIU Jun-dan[1,2,3]
(Mathematics & Information Science College, Hebei Normal University, Shijiazhuang 050024, China)[1]
(Hebei Key Laboratory of Computational Mathematics and Applications, Shijiazhuang 050024, China)[2]
(Institute of Mobile Internet of Things, Hebei Normal University, Shijiazhuang 050024, China)[3]

Abstract Considering the problems aroused by the traditional association rules mining algorithms which are lack of efficient data selection mechanism for users, especially not conducive to deal with multi-valued attribute data, this paper presented the redefinition and classification of multi-valued attribute data by using conceptual lattice, proposed an improvement of Apriori algorithm based on the KAF factor and the CHF factor to mine multi-valued attribute association rules and established a complete mining parameters adjustment mechanism acting very well in improving the speed and efficiency of mining, which is convenient for users to select key attribute values to mine and analyze rules while improving speed and mining algorithm efficiency. At the end of this paper, we illustrated the advantages of these new methods with the help of experimental data obtained from demographic data of a province, and the realistic application analysis and experimental results turn out that the improved mining algorithm has a better performance.

Keywords Multi-valued attribute, Concept lattice, Association rules, Apriori

1 引言

在数据挖掘研究领域中，关联规则(Association Rules)挖掘应用广泛，其作用是从海量数据集中发现属性间存在的、隐藏的、新颖的、有趣的关联或相关关系。然而，现有方法不能有效处理多值属性数据。作为数据分析的有力工具，概念格已经被人们应用到数据挖掘研究中。Wang等[1]给出了基于量化概念格的关联规则挖掘分析方法，运用概念节点展现规则可以避免组合爆炸问题，但该方法缺乏有效的多值属性数据处理机制，不利于用户选择针对性较强的数据进行分析。Bay Vo等[2]介绍了一种基于格和哈希表的关联规则挖掘方法，该方法引入了多种兴趣度度量值，但是无法满足用户动态分析频繁项集和规则的需求，也无法有效地处理多值属性项。Srikant和Agrawal在1996年首次提出多值关联规则挖掘问题及挖掘算法，将多值属性关联规则的挖掘转化为布尔型关联规则的挖掘[3]。Li等[4]介绍一种自适应的方法，以提高数字多值属性关联规则挖掘结果的离散分区。Prakash等[5]采用分离域的模糊规则集的一个定性方法进行多值属性关联规则挖掘，将数值属性转换成模糊属性来降低离群值的规则敏感性，但该方法计算复杂度较高，需要消耗更多的时间来产生各种离群规则集，增加了算法的时间复杂度。毛宇星等[6]提出了一种多层关联规则挖掘方法，用户通过设定控制阈值将

到稿日期：2013-05-27 返修日期：2013-07-22 本文受河北省科学技术研究与发展计划项目(072435158D, 09213515D, 09213575D)，河北师范大学硕士基金(201102002)资助。

郭晓波(1986－)，男，硕士生，CCF会员，主要研究方向为数据挖掘、智能信息处理，E-mail：xb_guo@163.com；赵书良(1967－)，男，博士，教授，博士生导师，主要研究方向为智能信息处理，E-mail：zhaoshuliang@sina.com(通信作者)；王长宾(1984－)，男，硕士生，主要研究方向为数据挖掘；赵娇娇(1986－)，女，硕士生，主要研究方向为自然语言处理、智能信息处理；刘军丹(1987－)，女，硕士生，主要研究方向为应用数学、智能信息处理。

相关性较高的项聚成一类，将聚类结果进行约简划分，节省了算法执行时间，但是该方法不便于用户灵活挑选不同的数据项或某些针对性较强的数据项进行关联规则挖掘来提高挖掘结果的质量，无法处理多值属性数据。Lee 等[7]采用多支持方式约束的方法从多值属性数据事务数据中挖掘关联规则，该方法利用模糊集概念将事务中的定量数据转换成基本项目，通过设置最小和最大支持度进行规则挖掘，但挖掘结果存在冗余。目前，国内外研究人员和学者对多值属性关联规则的研究内容主要集中于多值属性数据离散化、兴趣度量值分析等，大都存在以下不足[8-11]：用户无法挑选针对性较强的数据进行关联规则挖掘；所采用的算法不利于处理多值属性数据，挖掘速度和效率比较低；缺乏有效的挖掘参数调整机制；无法分析多值属性关联规则概念层间的关联模式。

概念格(Concept Lattice)将哲学的概念进行数学化的描述，实现了概念的一种形式化描述。概念格理论表达数据的基本形式是形式背景。在大量数据库应用中，对于数据的分析并非都是单值属性的形式背景-单值背景[12]，而更多是复杂的多值属性的形式背景-多值背景[13,14]。本文运用概念格将多值数据有机地组织起来，使数据之间的关系通过概念格节点的特化与例化关系生动简洁地体现出来，将关键属性因子 KAF(Key Attribute Factor)和概念层因子 CHF(Concept Hierarchy Factor)引入到 Apriori 算法中进行多值属性关联规则挖掘，方便用户选择关键数据进行分析，挖掘多值属性关联规则及不同概念层间的关联模式，提高挖掘速度和效率。结合概念格理论对多值属性数据进行了重新定义和分类，构建了完整的参数调整机制，使用户充分地参与到挖掘过程中，便于用户对频繁项集及关联规则进行动态分析和研究。文章利用概念格理论提出了多值属性关联规则挖掘方法，使用户可代替领域专家直接进行数据挖掘，在提高数据挖掘效率的同时，大幅提高了数据挖掘的可用性。

2 多值属性关联规则的概念格表示

2.1 项目集的概念格表示

概念格是一种基于二元关系的完备层次结构，它的每一节点称为一个概念。每个概念由外延和内涵两部分组成，概念的外延表示属于这个概念的所有对象的集合，而内涵则表示这些对象所共有的属性集合。作为数据分析的有力工具，概念格已经被人们广泛地应用于知识发现和数据挖掘领域。这里将项集与概念格相结合，研究概念格与频繁项目集之间的关系。

定义 1 属性又称为项，设 $A=\{a_1,a_2,\cdots,a_k,\cdots,a_n\}$(其中，$k\in N^+$，$1\leqslant k\leqslant n$，$a_k$ 称为一个项目，表示一个属性)为 n 个不同项目的集合。设定事务集 $T=\{(t_1,i_1),(t_2,i_2),\cdots,(t_k,i_k),\cdots,(t_m,i_m)\}$(其中，$k\in N^+$，$1\leqslant k\leqslant m$)，其中$(t_k,i_k)$代表一个概念，表示一个对象(事务)，$t_k$ 是对象(事务)的标识符，$i_k\subseteq A$ 是对象(事务)的属性值(项目)集合。

给定一个三元组(T,I,R)，将其称为形式背景(Formal Context)，其中 T 是事务的有限集合，I 是属性值的有限集合，R 是 $T\times I$ 上的二元关系，存在惟一的偏序集合与之对应，并且由这种偏序集合产生一种格结构，这种由形式背景(T,I,R)所诱导的格 L 称为一个概念格[12]。

概念格中的每个节点 $N=(T_k,I_k)$是一个二元组，其中 $T_k\in P(T)=2^T$，$I_k\in P(I)=2^I$，$P(T)$和 $P(I)$分别表示数据库中的事务集合和项目集合的幂集，对于给定的(T_k,I_k)定义如下映射：

$$f:2^T\to 2^I,\ f(T_k)=T_k'=\{i\mid i\in I,\forall t\in T_k\subset T,tRi\}$$

$$g:2^I\to 2^T,\ g(I_k)=I_k'=\{t\mid t\in T,\forall i\in I_k\subset I,tRi)\}$$

并且 $T_k'=I_k$，$I_k'=T_k$，则 T_k 和 I_k 分别称为概念的外延和内涵，其中 f 和 g 称为 T 的幂集和 I 的幂集之间的 Galois 连接[12]。

2.2 多值属性数据分类

多值背景[15,16]就是事务(记录)和属性之间不能仅仅用布尔型关系来表示，而是在原有的形式背景中出现了属性值的集合，并用具体的属性值来表示。许多事务涉及到多个属性值，而事务与属性之间的关系并不能简单地用“记录具有或者不具有某种属性”来描述。比如，在某省全员人口数据库中，“学历”、“文化程度”、“年龄”、“户口性质”等均称为多值背景，即事务与属性之间存在的关系无法只用“1”或“0”表示。为了便于挖掘任务的实现，本文提出适合多值属性关联规则可视化挖掘的多值背景定义，具体如下：

在多值属性集中，对于“年龄”、“世代间隔[1)]”等这样的表示数量化的属性项，其属性值都是用具体的数值来描述事务与属性之间的关系，则称该多值背景为数值型多值背景，其定义如下：

定义 2 设五元组(T,I,N,H_N,R_N)是一个数值型多值背景，其中 T 是事务集，I 是属性集，N 是数值型属性值的集合，H_N 是数值属性的概念层集合，$R_N\subseteq T\times I\times N\times H$ 是它们之间存在的一个四元关系，当且仅当对于任意 $t\in T$，$i\in I$，$h\in H$，有且只有一个 $n\in N$ 满足$(t,i,n,h)\in R_N$，用$(t,i,n,h)\in R_N$ 表示“对于属性 i，事务 t 在 h 上具有数值型属性 n”。若满足$(t,i,n_j,h_j)\in R_N$ 且$(t,i,n_j',h_j')\in R_N$，那么必有 $n_j=n_j'$，$h_j=h_j'$，其中 $j\in N^+$，表明 T 中同一个 I 的 N 在 H_N 上相等。

在实际应用中，很多属性项所具有的属性值为区间形式，即属性值都是以具体的区间值来描述事务与属性之间的关系，则称该多值背景为区间型多值背景，其定义如下：

定义 3 设五元组(T,I,S,H_S,R_S)是一个区间型多值背景，其中 T 是事务集，I 是属性集，S 是区间属性值的集合，H_S 是区间型的概念层集合，而 $R_S\subseteq T\times I\times S\times H$ 是表示它们之间存在的一个四元关系，当且仅当任意 $t\in T$，$i\in I$，$h\in H$，有且只有一个 $s\in S$ 满足$(t,i,s,h)\in R_S$，用$(t,i,s,h)\in R_S$ 表示对于属性 i，事务 t 在 h 具有区间型属性 s。若满足$(t,i,s_j,h_j)\in R_S$ 且$(t,i,s_j',h_j')\in R_S$，那必有 $s_j=s_j'$，$h_j=h_j'$，其中 $j\in N^+$。即 $s_j^L=s_j'^L$，$s_j^U=s_j'^U$，$h_j=h_j'$，其中 $s_j\in[s_j^L,s_j^U]$，$s_j'\in[s_j'^V,s_j'^U]$(s_j^L，s_j^U 表示 s_j 的最小、最大值，$s_j'^L$，$s_j'^U$ 表示 s_j'的最小、最大值)，表明 T 中同一个 I 的 S 在 H_S 上相等。

对于多值属性集中，如文化程度分为“高级”、“中级”和“初级”等，若其属性值都是以具体的类别值来描述事务与属性之间的关系，则称该多值背景为类别型多值背景，其定义如下：

1) 世代间隔是指人口增长率固定时母亲一代与女儿一代的间隔长度。马瀛通：《人口统计分析学》，北京：红旗出版社，第 696 页。

定义 4 设五元组(T,I,C,H_C,R_C)是一个类别型多值背景，其中 T 是事务集，I 是属性集，C 是类别型属性值的集合，H_C 是类别型的概念层集合，而 $R_C \subseteq T\times I\times C\times H$ 是它们之间存在的一个四元关系，当且仅当对于任意 $t\in T, i\in I, h\in H$，有且只有一个 $c\in C$ 满足$(t,i,c,h)\in I$，用$(t,i,c,h)\in R_C$表示“对于属性 i，事务 t 在 h 上具有类别型属性 c”。若满足$(t,i,c_j,h_j)\in R_C$ 且$(t,i,c_j',h_j')\in R_C$，那么必有 $c_j=c_j', h_j=h_j'$，其中 $j\in N^+$，表明 T 中同一个 I 的 C 在 H_C 上相等。

2.3 多值属性关联规则表示与求解

对于任意 $a\in I$，a 的取值可以为数值型、区间型和类别型。设 a 的取值集合为 V，若满足 $\forall v\in V_n$ 存在 $\upsilon,\mu\in N^+$，$\upsilon\leqslant\mu$使得 $v\in[\upsilon,\mu]$，则称 a_n 为数值型多值属性；若满足 $\forall v\in V_s$ 存在 $l,v\in N^+$使得 $v=[l,v]$，则称 a_s 为区间型多值属性；若满足 $\forall v\in V_c=[\alpha_1,\alpha_2,\cdots,\alpha_m]$，$m\in N^+$，则称 a_c 为类别型多值属性。

如果 a_n 为数值型属性值，$a_n=\langle a,\upsilon,\mu\rangle$（其中$\langle a,\upsilon,\mu\rangle\in I\times N^+\times N^+$），则三元组$\langle a,\upsilon,\mu\rangle$表示数值属性 a 的属性值在区间$[\upsilon,\mu]$上；如果 a_s 为区间型属性值 $a_s=\langle a,l,v\rangle$（其中$\langle a,l,v\rangle\in I\times N^+\times N^+$），则三元组$\langle a,l,v\rangle$表示数值属性 a 的属性值是$[l,v]$；如果 a_c 为类别型属性值，$a_c=\langle a,\alpha\rangle$（其中$\langle a,\alpha\rangle\in I\times N^+$}），则二元组$\langle a,\alpha\rangle$表示属性 a 的属性值为 α。由此可知，类别属性只与值相关，而数值或区间属性既可以与值相关联，也可以与区间相关联。元组$\langle a,\upsilon,\mu\rangle$、$\langle a,l,u\rangle$和$\langle a,\alpha\rangle$称为项(Item)，则 I 称为项集(ItemSet)。记为$\lhd i\rhd=\{x|\langle x,l,u\rangle\in i, i\subseteq I\}$，即$\lhd i\rhd$是项集 i 所包含的属性集合。

定义 5 若对于任意$\langle a,\upsilon,\mu\rangle$、$\langle a,l,u\rangle$和$\langle a,\alpha\rangle\in\lhd i\rhd$，存在$\langle a,q\rangle\in t_j$，使得 $\upsilon\leqslant q\leqslant\mu$ 或 $q=[l,v]$或 $q=\alpha$ 成立，则称事务 t_j 支持 i。

定义 6 多值属性关联规则是具有 $i_l\Rightarrow i_r$ 形式的蕴涵式，其中 $i_l,i_r\subset I$，并且$\lhd i_l\rhd\cap\lhd i_r\rhd=\emptyset$。如果 T 中有 $s\%$的事务支持 i_l 和 i_r，且 $c\%$的支持 i_l 的事务也支持 i_r，则该规则的支持度和置信度为 $s\%$和 $c\%$。

定义 7 不小于最小支持度阈值的项集称作频繁项集，即若 $support(L)\geqslant min\_sup$ 成立，那么称项集 $L\subseteq I$ 是频繁的，其中，项目集 $L\subseteq I$ 的支持度定义为：$support(L)=|L|/|T|$；同理，不小于最小支持度阈值的概念格中的概念称为频繁概念。

性质 1 对于概念 $C=(T,I)$，假设 $C_i=(T_i,I_i)$，$i\in N^+$，如果它是概念 C 的子概念，且 I_i 是频繁的，则 I 也是频繁的。

性质 2 对于概念 $C=(T,I)$，假设 $C_i=(T_i,I_i)$，$i\in N^+$，如果它是概念 C 的子概念，且 I 是不频繁的，则 I_i 也是不频繁的。

3 算法描述

传统的 Apriori 算法局限于设置最小支持度和最小置信度两个参数来进行关联规则挖掘，且仅能处理布尔类型的字段，无法有效地处理多值属性字段。针对这些问题，本文对 Apriori 算法进行了改进，引入关键属性因子(Key Attribute Factor，KAF)和概念层因子(Concept Hierarchy Factor，CHF)进行多值属性关联规则挖掘，方便用户进行有选择性的分析和挖掘，较大地提升了挖掘速度和效率。

本节主要介绍了频繁项集、关联规则挖掘算法。首先，对引进 KAF 和 CHF 因子的 Apriori 改进算法进行了详细描述，其中 genFreqItemset()用于挖掘多值属性数据之间的频繁模式-频繁项集，并对频繁项集进行初始化；然后，对关联规则挖掘算法 genRule()进行了介绍，其主要功能是挖掘多值属性数据的关联模式-关联规则。

3.1 频繁项集挖掘算法

算法采用与 Apriori 算法类似的逐层探索迭代方法，k 项集用于探索$(k+1)$项集。为了提高挖掘效率和执行速度，本文引入了 KAF 因子和 CHF 因子进行有选择性的频繁项集挖掘。在初始情况下，通过设置 KAF 和 CHF 值的大小，查询由关键因子构成的数据集，在以后的 k 频繁项挖掘中，利用 CHF 因子对其进行初始化，将不同的频繁项集划分到不同的抽象层，便于进行可视化展示和各种应用分析。算法 1 为频繁项集挖掘算法。

算法 1 genFreqItemset()

```
输入：数据集 DB，最小支持度 minSup，KAF 因子和 CHF 因子。
输出：频繁项集 L。
genFreqItemset(DB,minSup,KAF,CHF)
(1)TID=get_key_value(DB,KAF,CHF);//TID 是根据 KAF 和 CHF 因子选择的关键数据集
(2)L1=get_frenqent_1_itemset(TID);
(3)freqInitial(频繁 1-项集,CHF);//初始化 1-itemset
(4)FOR (k=2;Lk-1≠Φ;k++) DO {
(5)     Ck=genCandidate(Lk-1,CHF);//根据 CHF 因子值对 k-item 进行概念分层
(6)     FOREACH 每个事务 t∈TID DO{
(7)       FOREACH 每个候选项 c∈Ck DO{
(8)         IF c∈事务 t THEN
            c.Count++;//支持度计数增值
(9)       }//end foreach
(10)    Lk={c∈Ck|c.count>=minSup};
(11)} //end for
(12)return Sort(L=∪Lk);//将频繁项集按包含项的个数进行排序
```

genFreqItemset()算法是在多值属性值频繁项集挖掘情况下对 Apriori 算法的改进，引入了确定关键数据集和频繁项概念分层的生成方法，提高了算法运行速度和挖掘结果的质量。

genFreqItemset()算法的优点是利用 $KAF\{N_i,S_i,C_i\}$和 $CHF\{N_j,S_j,C_j\}$因子，设置数值型属性值集合 N_i、区间型属性值集合 S_i 和类别型属性的集合C_i，并将不同的频繁项集划分到相应的概念层$\{N_j,S_j,C_j\}$上，该方法具有较强的可选择性和灵活性。由于引进关键属性因子和概念层因子来定义查询的数据集，使得产生冗余项集的问题在该类算法中也得到了很好的解决。

3.2 关联规则挖掘算法

算法对关联规则挖掘进行概念分层处理，从而有效减少了规则的搜索时间。算法 2 为关联规则挖掘算法。

算法 2 genRule()

```
输入：频繁项集 L，最小置信度 minConf，CHF 因子。
输出：关联规则 Rule。
genRule(L,minConf,CHF)
(1)FOR 每一个频繁 k-项集 fk,k>=2 DO{
(2)  Ct=getSubset(fk);//取每个频繁的子集
(3)  IF Ct.Count > 0 THEN
```

```
(4)    FOREACH 每个频繁子集 c∈Ct DO{
(5)    Conf=TID. FindSupport(fk) /TID. FindSupport(c);
       //取当前频繁子集的支持度
(6)    IF Conf >=minConf THEN
          Rules(频繁子集→Ct. Remove(频繁子集 c),CHF);//初
          始化规则前件和后件
(7)    }//end foreach
(8)  }//end if
(9)}//end for
(10)return RuleSet;//返回规则集合
```

genRule()算法是在多值属性值关联规则挖掘情况下对Apriori算法的改进，增加了确定关联规则概念分层的生成方法。其优点是运用CHF因子将规则的前件和后件进行概念层初始化，便于用户分析与研究不同概念层次的规则信息。

4 实现过程

关联规则挖掘的主要目的就是运用关联规则挖掘算法，通过设置相关参数从海量数据集中发现隐藏的、新颖的、属性间存在的有趣的关联或相关关系，如Apriori算法，通过确定最小支持度和置信度来对数据进行关联规则挖掘，研究数据之间有趣的关系模式。然而一个重要的问题是，在关联规则挖掘过程中，确定合适的输入参数往往比较困难，不便于动态分析关联规则挖掘过程中的相关操作。现有文献中的关联规则挖掘缺少统一的参数调整机制，导致用户无法通过设置其它参数来发现感兴趣的信息，例如关键字段的选择。因此，建立完整统一的参数调整机制是非常重要的，它不仅可以帮助用户在整个挖掘过程中动态地进行参数调整，而且在很大程度上提高了挖掘速度和效率。针对上述问题和多值规则挖掘的特点，结合概念格的多值背景理论，建立了以支持度(Sup)、置信度(Conf)、关键属性因子(Key Attribute Factor，KAF)和概念层因子(Concept Hierarchy Factor，CHF)为基础的参数调整机制，在整个挖掘过程中通过调整Sup、Conf、KAF和CHF参数的大小来挖掘相应的频繁项集和关联规则，较大地提高了挖掘过程的灵活性，让用户完全参与到挖掘的过程中，如图1所示。通过参数模块Pas设置KAF因子和CHF因子值从输入的数据集DB中选择关键属性值和不同的概念层，用户可以交互地调整KAF和CHF动态分析数据，最后设置Sup和Conf值挖掘频繁项集和关联规则。具体介绍如下：

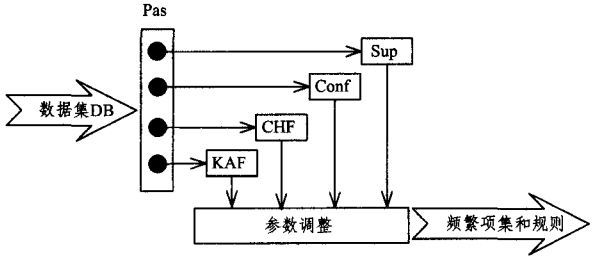

图1 参数调整机制

1)多值属性值离散化，根据实际情况在关联规则挖掘之前对不同形式背景中多值属性字段进行离散化处理，依据定义2—定义4对记录中的多值属性字段进行转换。本文选择全员人口数据库中的4个属性值(年龄、户口性质、世代间隔、管理地)(如表1所列)，对数值型字段“年龄”和“世代间隔”、类别型字段“户口性质”和“管理地”进行离散化操作，得到如表2—表5所列的结果。离散化完成后，结果如表6所列。

表1 人口数据表

人口编号	年龄	户口	世代间隔	管理地
1307001	22	其它	[15—19]	农村
1307002	28	非农业	[25—29]	城镇
1307003	24	农业	[20—24]	农村

表2 年龄属性

年龄	整数值
[20—24]	1
[25—29]	2
[30—34]	3

表3 户口性质属性

户口	整数值
农业	1
非农业	2
其它	3

表4 世代间隔属性

间隔	整数值
[15—19]	1
[20—24]	2
[25—29]	3

表5 管理地属性

管理地	整数值
农村	0
城镇	1

表6 人口数据表(离散后)

人口编号	年龄	户口	世代间隔	管理地
1307001	1	3	1	0
1307002	2	2	3	1
1307003	1	1	2	0

2)设置KAF因子，主要是选择数据库记录中的关键字段。关联规则挖掘的一个重要任务就是生成频繁项集，当数据项较多时，生成的项集和规则的数量也会非常大，而其中包含许多用户不感兴趣的项集。直接挖掘有趣的规则，最关键的一点是如何为算法选择适当的属性值，因此在关联规则挖掘过程中，用户通过设置$KAF\{N_{ki},S_{ki},C_{ki}\}$因子来选择记录(项目)集合$A=\{a_1,a_2,a_3,\cdots,a_k,\cdots,a_n\}$中的关键属性项进行分析(其中，$k$表示关键属性，$i$表示集合项数)，其中$N_{ki}$表示数值型属性值集合，$S_{ki}$表示区间型属性值集合，$C_{ki}$表示类别型属性的集合，其中$i\in N^+$。如：$KAF\{$(年龄)，(间隔)，(户口性质)$\}$。

3)设定CHF因子，主要是将数据库中的字段项按$CHF\{N_{hj},S_{hj},C_{hj}\}$因子值(其中，$h$表示概念层，用于区分关键属性因子，$j$表示属性所属层级)进行概念分层初始化，其中$N_{hj}$表示数值型属性值的概念层数，$S_{hj}$表示区间型属性值的概念层数，$C_{hj}$表示类别型属性的概念层数，其中$j\in N^+$，如$CHF\{3,3,3\}$。事实上，关系表中存在如下两类概念：一类是事务概念，每个事务概念是事务集的某个子集；另一类是属性值概念，每个属性值概念是属性取值域的某个子集。通常，每个属性a的取值中都存在一个概念层H。这些概念层次是蕴涵在属性取值中的重要信息。如图2概念分层所示，其为由$\{\{t_1$，文化$\}$，$\{t_2$，初级$\}$，$\{t_3$，小学$\}$，$\cdots\}$等构成的概念层形式，其中$\{\{t_2$，初级$\}$，$\{t_4$，中级$\}$，$\{t_5$，高级$\}\}$具有相同(第二层)的层关系；$\{\{t_6$，小学$\}$，$\cdots$，$\{t_9$，中专$\}$，$\cdots$，$\{t_{12}$，本科$\}\}$具有相同(第三层)的层关系。

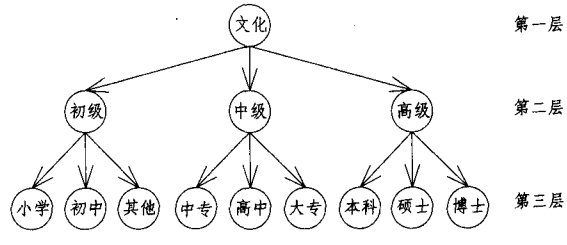

图2 概念分层

4)完成前3步操作后,设定最小支持度Sup、最小置信度Conf的大小进行多值关联规则挖掘。

5 实验结果分析

为了验证本文改进算法的正确性、高效性和其它各种性能,本文采用以下实验环境:测试程序在Windows Server 2008系统上运行,主频3.10GHz,4GB RAM,VS2008,数据库系统为ORACLE 10G,算法实现均采用C#语言,测试数据为某省全员人口数据。经过离散化处理后,所生成数据集的具体控制参数含义及其缺省值详如表7所列。

表7 测试数据的相关参数

参数符号	具体含义	设置大小
|TID|	记录个数	100～324000
I	项的数目	5～15
A	项的属性个数	1～5
F	频繁项集个数	300
V	最大频繁项集平均长度	2～10
H	概念分层数	3～4

本文主要从记录数量变化、支持度大小变化和项数目变化3个方面对Apriori算法和改进算法进行了具体的实验比较和分析,具体结果如图3—图5所示。

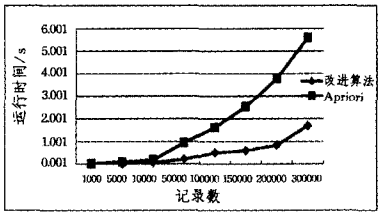

图3 运行时间与记录数量

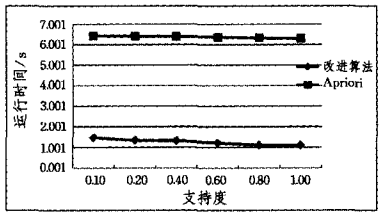

图4 运行时间与最小支持度

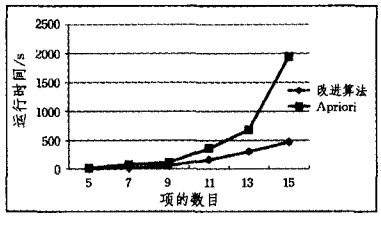

图5 运行时间与项的数目

1)运行时间与记录数

当同时取最小支持度5.0时,项目数量为10个,记录数从1000条逐步增至300000条,采用Apriori算法与改进算法对上述全员人口数进行性能测试,由图3可以看出,在同等条件下,改进算法的执行速度明显快于Apriori算法。在记录数量相对较小的情况下,二者区分效果不够明显,但是随着数据的不断增加,二者的执行速度相差效果逐步明显,当记录数为300k时,改进算法执行时间为1.684s,只有Apriori算法执行时间5.606s的30%,性能提高了3.3倍。因此,引入KAF和CHF因子的Apriori改进算法进行多值属性关联规则挖掘,较大地提升了挖掘速度和效率。

2)运行时间与支持度

当取记录数为324000条时,最小支持度从0.1逐步增至1.0,分别利用Apriori算法和本文算法对上述数据进行测试。由图4可以看出,在数据记录数相同的情况下,随着最小支持度的不断增加,两种算法的运行时间减小,但在整个变化过程中改进算法的平均执行时间为1.25s,而Apriori的平均执行时间为6.3s,性能约提高了5倍。

3)运行时间与项数目

当取记录数为324000条时,最小支持度为5.0,项的个数从5个逐步增至15,分别对两种算法进行测试。由图5的实验结果可知,在项的数目相对较少的情况下,二者区分效果不够明显,但是伴随数据项数量的不断增加,改进算法的执行效率显然比Apriori算法具有明显优势,当数据项增至15个时,Apriori算法的执行时间是1945.987s,而改进算法执行时间是469.987s,性能比提高了4倍多。实验结果表明,改进算法在大规模数据项目的数据库挖掘中具有更好的性能表现。

以上实验结果表明,在多值属性关联规则挖掘中,在相同条件下,引入KAF和CHF因子的Apriori算法的执行速度比Apriori算法快,显著地提高了挖掘性能。通过设置KAF和CHF因子,方便用户进行有选择性的数据分析与挖掘,提升了挖掘速度和效率,具有较强的实用性。

结束语 本文通过对多值属性数据的分析与研究,提出一种新的基于概念格的多值属性关联规则挖掘方法,采用基于KAF因子和CHF因子的Apriori改进算法进行多值属性关联规则挖掘,执行速度与Apriori算法相比有较大提高,具有更好的挖掘效率和性能。针对数据库中存在大量多值属性字段,运用概念格理论对其进行了分类,为多值数据进行离散化提供了理论依据;在关联规则挖掘过程中建立了完整的挖掘参数调整机制,极大地提升了挖掘速度和效率;详细介绍了具体的频繁项集、关联规则挖掘算法及关联规则挖掘过程,并对具体算法和挖掘过程进行了分析。最后,通过运用某省全员人口数据对算法进行了具体实现和分析,实验结果表明改进算法具有更好的挖掘效率和性能。本文提出的改进算法与Apriori算法类似,由于数据存储于数据库中,需要多次读取数据库,存在较大时间开销。在下一步的工作中,我们将研究如何把挖掘出来的频繁项集和关联规则进行可视化展示;如何利用频繁项集和关联规则中所含数据项之间的语义联系与应用背景,把频繁项集和规则转换为领域知识进行可视化展示。

参 考 文 献

[1] Wang D X, Xie Q. Analysis of association rule mining on quantitative concept lattice [C]// Artificial Intelligence and Computational Intelligence, LNCS7530. Berlin: Springer-Verlag, 2012: 142-149

[2] Bay Vo, Bac Le. Interestingness measures for association rules: combination between lattice and hash tables [J]. Expert Systems with Applications, 2011, 38(9): 11630-11640

[3] Srikant R, Agrawal R. Mining quantitative association rules in large relational table [C]// SIGMOD'96 Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. New York: ACM 1996: 1-12

(下转第309页)

4.3.3 与其它 SVM 模型的识别性能对比

采用 AFSA、PSO、GA 对 SVM 的参数进行优化，建立相应的人体运动姿态识别模型，并对测试样本进行检测，得到的人体运动姿态平均识别正确率如图 7 所示，各模型的参数优化过程如图 8 所示。

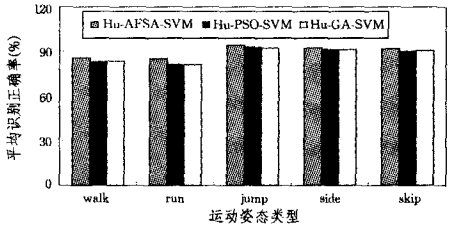

图 7 各种 SVM 模型的识别性能对比

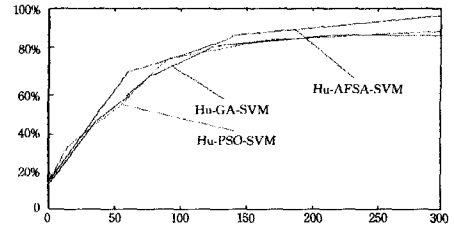

图 8 RBF 神经网络的训练误差曲线

由图 7 可知，Hu-AFSA-SVM 模型的人体运动姿态平均识别正确率要高于对比模型 HU-GA-SVM 和 HU-PSO-SVM，说明采用 AFSA 对 SVM 参数进行优化，可以获得比 GA 和 PSO 算法更优的 SVM 参数，这样进一步提高了人体运动姿态识别效果。同时从图 8 可知，Hu-AFSA-SVM 收敛明显加快，在 120 左右时，基本达到了 HU-GA-SVM 和 HU-PSO-SVM 最高识别正确率，对比结果表明，Hu-AFSA-SVM 是一种识别正确率高、速度快的人体运动姿态识别模型。

结束语 将 Hu 不变矩和人工鱼群算法优化支持向量机应用于人体运动姿态识别，并采用仿真实验对识别性能进行了验证。采用人工鱼群算法优化 SVM 的参数，有效克服了核函数参数选择对 SVM 识别效果的影响，采用 Hu 不变矩作为人体运动姿态识别特征，可以获得更优的人体运动姿态识别正确率。仿真结果表明，Hu-AFSA-SVM 提出的人体运动姿态识别方法是行之有效的。

参 考 文 献

[1] 凌志刚，赵春晖，梁彦，等. 基于视觉的人行为理解综述[J]. 计算机学报，2008，25(9)：2570-2578

[2] 谷军霞，丁晓青，王进生. 行为分析算法综述[J]. 中国图像图形学报，2009，19(3)：377-387

[3] 朱强，庄越挺，陈家实，等. 基于色块的人体运动跟踪[J]. 模式识别与人工智能，2001，14(4)：486-492

[4] 黄国范，程小平，任菲. 人体动作姿态的自动识别方法研究[J]. 西南师范大学学报：自然科学版，2010，35(4)：136-140

[5] 李宗民，刘玉杰，李振波，等. Bezier 矩及其在人体姿态识别中的应用[J]. 计算机工程与应用，2005，24(5)：38-40

[6] 李宁，须德，傅晓英. 结合人体运动特征的行为识别[J]. 北京交通大学学报，2009，33(2)：6-16

[7] 朱望飞，马义德，邱秀清. 基于 PCNN 的高斯混合模型运动检测改进方法[J]. 兰州大学学报：自然科学版，2009，45(2)：131-137

[8] 高晶敏，梁菁菁，李春云. 基于 RBF 神经网络的人体动态姿态识别算法[J]. 北京信息科技大学学报，2011，26(4)：27-29

[9] 俞洋，殷志锋，田亚菲. 基于自适应人工鱼群算法的多用户检测器[J]. 电子与信息学报，2007，29(1)：121-124

[10] 向昌盛，周子英，张林峰. 基于均匀设计的最小二乘支持向量机改进算法[J]. 计算机仿真，2011，28(3)：194-197

[11] 李红波，向南，吴渝. 一种改进的室内运动人体检测与轮廓提取算法[J]. 重庆邮电大学学报：自然科学版，2008，20(2)：225-229

(上接第 271 页)

[4] Li D C, Zhang M. A new approach of self-adaptive discretization to enhance the Apriori quantitative association rule mining [C]// ISDEA'12: Proceedings of the 2012 Second International Conference on Intelligent System Design and Engineering Application. Washington, DC: IEEE Computer Society, 2012: 44-47

[5] Prakash S, Parvathi R. Qualitative approach for quantitative association rule mining using fuzzy rule set [J]. Journal of Computational Information Systems, 2011, 17(6): 1879-1885

[6] Mao Yu-xing, Chen Tong-bin, Shi Bo-le. Efficient method for mining multiple-Level and generalized association rules [J]. Journal of Software, 2011, 22(12): 2965-2980

[7] Lee Y Y C, Hong T P, Chen C H. Mining generalized association rules with quantitative data under multiple support constraints [C]// ICCCI'10: Proceedings of the Second international conference on Computational collective intelligence: technologies and applications. Berlin: Springer Verlag, 2010: 224-231

[8] Alatas B, Akin E, Karci A. MODENAR: multi-objective differential evolution algorithm for mining numeric association rules [J]. Applied Soft Computing, 2008, 8(1): 646-656

[9] Pachón álvarez V, Mata Vázquez J. An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization [J]. Export Systems with Applications, 2012, 39(1): 585-593

[10] Martínez - Ballesteros M, Riquelme J. Analysis of measures of quantitative association rules [C]// HAIS'11: Proceedings of the 6th international conference on Hybrid artificial intelligent systems. Berlin: Springer Verlag, 2011(6679): 319-326

[11] Shaharanee I N M, Hadzic F, Dillon T S. Interestingness measures for association rules based on statistical validity [J]. Knowledge-Based Systems, 2011, 24(3): 386-392

[12] Ganter B, Wille R. Formal concept analysis: mathematical foundations [M]. Berlin: Springer Verlag, 1999

[13] Gugisch R. Many-valued context analysis using descriptions [C]// ICCS2001: Conceptual Structures: Broadening the Base Conceptual Structures: Broadening the Base Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2001: 157-168

[14] Nguyen T T, Hui S C, Chang K. A lattice-based approach for mathematical search using formal concept analysis [J]. Expert Systems with Applications, 2012, 39(5): 5820-5828

[15] Ourida B B S, Wafa T. Formal concept analysis based association rules extraction [J]. International Journal of Computer Science Issues, 2011, 8(4): 490-497

[16] Gély A, Medina R, Nourine L. Representing lattices using many-valued relations [J]. Information Sciences, 2009, 179 (16): 2729-2739