基于多维标度的快速挖掘关联规则算法*

肖 利 金远平 徐宏炳 王能斌

(东南大学计算机科学与工程系 南京 210096)

E-mail: lxiao@seu.edu.cn

摘要 挖掘关联规则是数据挖掘研究的一个重要方面.文章在分析其基本模型和研究多维标度基本性质的基础上,提出一个新的基于多维标度的挖掘关联规则算法.该算法以数据项间的关联度量为依据,将各个数据项投影到多维空间上,进行降维处理,最后将数据项集间的关联关系以可视结果提供给用户.

关键词 数据挖掘,知识发现,关联规则,多维标度法,可视.

中图法分类号 TP311

数据挖掘(data mining)是一个从数据中析取潜在有用的、先前未知的和最终可理解的知识的过程.挖掘关联规则是数据挖掘研究的一个重要方面,关联规则描述的问题是:在给定交易(transaction)数据库中,每个交易对应于一个数据项集,关联发现函数作用在这个交易数据库上,返回各项集间存在的密切关系.自从1993年以来,数据挖掘领域的研究者在挖掘关联规则上做了大量的工作,使之成为一个具有普遍和实用意义的数据挖掘技术,其中主要的工作有:经典的先验(Apriori)算法[1]及其扩展[2~4],广义关联规则[5,6]、量化关联规则[7]的挖掘算法以及增量挖掘关联规则算法[8]等.

挖掘关联规则算法是从大量的数据中挖掘关联规则,因此产生的关联规则的集合也非常大,这又将产生一个新的知识管理问题.而且数据挖掘系统的用户往往可能只对部分数据项集的关联规则有兴趣,并希望在较短的时间内得到有关数据项集间是否具有关联关系的提示.以先验算法为代表的挖掘关联规则算法需要对数据库进行多次扫描,反复迭代直至产生数据项集的全部关联规则为止.

本文提出的基于多维标度的快速挖掘关联规则算法,以数据项间的关联度量为依据,将各个数据项投影在多维空间上,然后进行降维处理,最后将数据项集间的关联关系以可视结果提供给用户.用户可以根据可视结果的提示,选择自己感兴趣的并且具有关联关系的数据项集,进行定向挖掘[4].这样就大大减少了扫描数据库的次数,有较好的实用意义.

1 关联规则基本模型

设$I=\{i_1,i_2,\ldots,i_m\}$是一个数据项集,D是一个交易集,每条交易T对应于一个数据项子集,即$T\subseteq I$.每条交易由一个TID(transaction identifier)标识.对数据项集X,当且仅当$X\subseteq T$,称交易T包含X.关联规则是形如$X\Rightarrow Y$的蕴含式,其中$X\subset I$,$Y\subset I$且$X\cap Y=\varnothing$.交易集D中的规则$X\Rightarrow Y$由置信度c(confidence)和支持度s(support)约束.置信度c定义为D中含有X中的交易的$c\%$,也含有Y.支持度s定义为:包含$X\cup Y$的交易占D的$s\%$.置信度表示蕴含式的强度,支持度表示在规则中出现该型式的频度,具有高置信度和高支持度的关联规则称为强关联规则.挖掘关联规则就是在大型数据库中发现强关联规则.该任务可以分解为以下两步:

---

* 本文研究得到新疆教委课题和江苏省自然科学基金资助.作者肖利,1968年生,博士生,主要研究领域为数据挖掘,数据仓库.金远平,1957年生,教授,主要研究领域为数据库系统.徐宏炳,1948年生,教授,主要研究领域为数据库系统.王能斌,1929年生,教授,博士生导师,主要研究领域为数据库,信息系统.

本文通讯联系人:肖利,南京 210096,东南大学计算机科学与工程系

本文1997-10-21收到原稿,1998-08-17收到修改稿

(1) 求数据项频集的集合. 数据项频集(frequent set)是支持度大于预定义的最小支持度的数据项子集.

(2) 使用数据项频集产生关联规则.

可以看出,在以上两步中最重要的是第1步,如果已经得到数据项频集,可以直接导出对应的关联规则[1]. 因此,挖掘关联规则的算法的工作主要集中在如何高效地发现数据项频集.

2 求解数据关联问题的多维标度法

在多维标度法中,我们将各个数据项投影到多维坐标中,进行降维处理,得到在低维坐标中数据项间的关联关系的直观描述. 因为一维标度是多维标度法的基础,所以,我们首先来介绍一维标度法.

2.1 一维标度法

设 $I=\{a_1,a_2,\ldots,a_n\}$ 是一个数据项集,假定数据项名 a_i 为数字型. D 是一个交易集,$s_{a_i,a_j}(a_i\neq a_j)$ 为数据项 a_i,a_j 之间的支持度,即在交易集 D 中,同时包含数据项 a_i,a_j 的交易的个数在 D 中的比率,显然 s_{a_i,a_j} 为正值. 我们希望对每个数据项 a_i 确定一个实数 x_{a_i},以 x_{a_i} 为坐标将数据项 a_i 表示为一维空间中的一个点. 选择坐标 x_{a_i} 的原则是:对于任意两个数据项 a_i,a_j 所对应的 x_{a_i},x_{a_j},点 x_{a_i} 和点 x_{a_j} 之间的距离 $|x_{a_i}-x_{a_j}|$ 应能反映数据项 a_i,a_j 之间的支持度 s_{a_i,a_j} 的大小. 距离越长,支持度越小,关联性就越弱. 为简单化,下面我们将 s_{a_i,a_j} 简写为 s_{ij},将 x_{a_i} 简写为 x_i.

例如,对于数据项1、2、3,假定支持度 $s_{1,2}=7,s_{1,3}=2$,则距离 $|x_1-x_2|$ 小于距离 $|x_1-x_3|$.

为达到此目的,使用以下准则:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=0, \tag{1}$$

$$\sum_{i=1}^{n}x_i^2=1. \tag{2}$$

在上述条件(1)、(2)之下,选择 $x_i(i=1,2,\ldots,n)$,使得 $Q=\sum_{i,j=1}^{n}\sum_{i\neq j}(-s_{ij})(x_i-x_j)^2$ 达到最大. 条件(1)是取 n 个坐标点的重心为坐标原点,条件(2)相当于坐标的有限尺度化.

将 Q 表示成二次型的形式:

$$Q=\sum_{i=1}^{n}(-\sum_{i\neq j}(s_{ij}+s_{ji}))x_j^2+2\sum_{i,j=1}^{n}\sum_{i\neq j}(s_{ij}+s_{ji})x_ix_j.$$

因为支持度 s_{ij} 具有对称性,则

$$Q=2(\sum_{i=1}^{n}(-\sum_{i\neq j}s_{ij})x_i^2+2\sum_{i,j=1}^{n}\sum_{i\neq j}s_{ij}x_ix_j).$$

对于上式,可以构造一个 $n\times n$ 阶对称矩阵 $S_{n\times n}=(s_{ij})$,其中 $s_{ij}=-\sum_{i\neq j}s_{ij},i,j=1,2,\ldots,n$. 这样,$Q$ 的条件最大值实际上就是在条件(1)、(2)之下求由矩阵 S 生成的二次型的最大值,我们把问题最终归结为求解特征方程 $S\bar{x}=\lambda\bar{x}$ 的特征值和特征向量. 所求的一维标度 $\bar{x}=(x_1,x_2,\ldots,x_n)$ 是 S 的最大特征值所对应的单位特征向量.

2.2 多维标度法

在挖掘关联规则的实际应用中,当数据项的数量大时,用一维空间不易充分反映数据项之间的关联性,有必要考虑多维空间的标度问题. 寻求如下的 $n\times p$ 阶矩阵,

$$X_{n\times p}=\begin{bmatrix}x_{11}&x_{12}&\ldots&x_{1p}\\x_{21}&x_{22}&\ldots&x_{2p}\\\ldots&\ldots&\ldots&\ldots\\x_{n1}&x_{n2}&\ldots&x_{np}\end{bmatrix}=\begin{bmatrix}\bar{x}'_{(1)}\\\bar{x}'_{(2)}\\\ldots\\\bar{x}'_{(n)}\end{bmatrix}=(\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n),$$

其中的行向量 $\bar{x}_{(i)}$ 表示第 i 个标度点的坐标($i=1,2,\ldots,n$),列向量表示 n 个标度点的第 k 个坐标($k=1,2,\ldots,p$). 同理,仍要求 X 中的元素满足条件:

$$\vec{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik} = \frac{1}{n}\vec{1}'\vec{x}_k = 0, \qquad k = 1,2,\ldots,n, \tag{3}$$

即要求 n 个数据项的标度是中心化的. 此外，还要求 X 中的各列是正交，

$$\vec{x}_k\vec{x}_l = \delta_{kl} = \begin{cases} 1, & k = l \\ 0, & k \neq 1 \end{cases} \qquad k,l = 1,2,\ldots,n. \tag{4}$$

各个标度点间在 P 维空间中的距离平方为

$$d_{ij}^2 = \| \vec{x}_{(i)} - \vec{x}_{(j)} \|^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2, \qquad i,j = 1,2,\ldots,n.$$

为使这些距离与关联度量 $-s_{ij}$ 在顺序上尽可能一致，把问题归结为在条件(3)、(4)之下使表达式：

$$Q = -\sum_{i,j=1}^{n}\sum_{i\neq j} s_{ij}d_{ij}^2 = -\sum_{k=1}^{p}\sum_{i,j=1}^{n}\sum_{i\neq j} s_{ij}(x_{ik} - x_{jk})^2 = \sum_{k=1}^{p}\vec{x}'_k S\vec{x}_k$$

取最大值 S. 的前 p 个最大特征值所对应的正规直交特征向量是我们需要的解 $\vec{x}_k(k=1,2,\ldots,p)$，由它们所构成的 $n\times p$ 阶矩阵 $X=(\vec{x}_1,\vec{x}_2,\ldots,\vec{x}_p)$ 的各行给出了 n 个点在 P 维空间中的坐标.

3　基于多维标度的挖掘关联规则算法

3.1　特征方程的性质

特征方程 $S\vec{x}=\lambda\vec{x}$ 具有以下几条性质.

性质 1. 由于 S 是对称的，它的所有的特征值都是实的，按大小顺序排列，可记为 $\lambda_1\geqslant\lambda_2\geqslant\ldots\geqslant\lambda_n$.

性质 2. 如果所有的 $s_{ij}\leqslant 0$，则 S 必为非负定的，从而其所有的特征值非负.

性质 3. 对关联度量 s_{ij} 作“加常数”变换：

$$\tilde{s}_{ij} = s_{ij} + c,$$

$$\tilde{s}_{ii} = -\sum_{i\neq j}\tilde{s}_{ij} = s_{ii} + c - cn,$$

其中 $i\neq j, i,j=1,2,\ldots,n, c$ 为任意常数. 结果相应矩阵 S 的特征向量不变，只是特征值与原来的 λ 相差一个常数 cn.

基于以上性质，在实际计算中，支持度 s_{ij} 为正值，所以我们要选取恰当的常数 $c<0$，对 s_{ij} 作加常数变换后，使所有的 $\tilde{s}_{ij}\leqslant 0$，这样可以保证 $\tilde{S}=(\tilde{s}_{ij})$ 非负定，从而所有的特征值皆为非负. 这样选取的最大特征值就不会为 0，所得到的解也就满足中心化条件(3).

3.2　模板算法

基于多维标度的挖掘关联规则算法如下所示.

(1) $c=getconst(\ )$；//确定常数 c，一般取为 -1；

(2) $L_1=\{frequent\ 1-itemsets\}$；

(3) $C_2=apriori\text{-}gen(L_1)$；

(4) $L_2=getfreItems(C_2)$；

(5) $S=build\text{-}change(L_2)$；

(6) $get\text{-}eig(S,x,d)$；

(7) $X=subselect(x,d,p)$；

(8) $plot(X)$.

L_k 是 k 位数据项的数据项频集，其数据结构为：频集中的每个成员有两个域：(1) 数据项集 $itemset$；(2) 支持数 $count$. 例如，$L_2.itemset=\{\{2\ 3\},\{4\ 5\},\{4\ 7\}\}$，$L_2.count=\{5,4,9\}$，即数据项 2 与 3、数据项 4 与 5、数据项 4 与 7 之间的支持数分别为 5、4、9. C_k 为 k 位数据项的数据项频集的候选集，其数据结构与 L_k 相同.

我们首先扫描数据库，生成1位数据项频集 L_1，利用函数 *apriori-gen*(L_1)* 由 L_1 产生候选集 C_2，函数 *getfreItems*(C_2)的功能是从 C_2 选取支持度大于预定义的最小支持度的数据项集 L_2.

函数 *build-change*(L_2)的功能是利用 L_2 构造关联性矩阵 S，然后将其变换为非负定矩阵 $\tilde{S}$，具体步骤为：

(1) $s_{ij}=change(L_2)$, $i\neq j$;

(2) $\tilde{s}_{ij}=s_{ij}/\|D\|-c$, $i\neq j$;

(3) $\tilde{s}_{ij}=-\sum_{i\neq j}\tilde{s}_{ij}$, $i,j=1,2,\ldots,sum(L_2)$.

其中 $\|D\|$ 为数据库 D 中 TID 的个数，$sum(L_2)$表示数据项频集 L_2 中数据项的个数. 函数 *change*(L_2)以数据项频集 L_2 构造关联性矩阵 S，例如，如果 $L_2.itemset=\{\{2\ 3\},\{4\ 5\},\{4\ 7\}\}$，$L_2.count=\{5,4,9\}$，那么相应的 $s_{23}=5$，$s_{45}=4$，$s_{47}=9$.

函数 *get-eig*(S,x,d)求矩阵 S 的特征向量 x 和特征值 d，函数 *subselect*(x,d,p)的功能是从特征值 d 中选出前 p 个最大特征值所对应的特征向量 $\tilde{x}_1,\tilde{x}_2,\ldots,\tilde{x}_p$，因为要进行二维可视化，我们选择 $p=2$，即前两个最大特征值所对应的特征向量 $\tilde{x}_1,\tilde{x}_2$.

矩阵 $X=(\tilde{x}_1,\tilde{x}_2)$的 n 行即为所求的各数据项点在2维空间的标度，函数 *plot*(X)将其可视化.

3.3 举例说明

应用第3.2节中的算法在交易数据库中挖掘关联规则.

首先，在数据库中挖掘数据项频集 L_1 和 L_2(假设最小支持数为1)，利用 L_2 构造矩阵 S，从矩阵 S 中看到：数据项集{2 3}的支持数为5，数据项集{4 5}的支持数为4，等等，最后的可视结果如图1所示. 分析图1，我们可得到如下提示：(1) 数据项5与6，数据项7、3与4，数据项10与11之间距离较近，因此数据项集{5 6}、{3 4 7}、{10 11}最有可能是数据项频集；(2) 数据项2、8、9，数据项8、11之间距离较远，随着最小支持度的增大，数据项集{2 8 9}、{8 11}不可能是数据项频集；(3) 数据项5与6可能是整个数据项集的重心，进行定向挖掘[4]时需要注意.

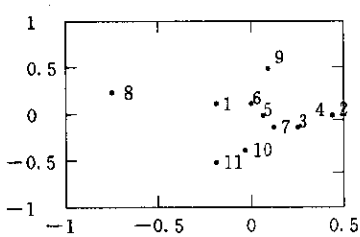

图1

$$S=\begin{bmatrix}
0&5&5&2&9&5&4&9&4&3&5\\
5&0&5&7&4&5&4&1&4&3&3\\
5&5&0&5&4&8&4&1&4&3&7\\
2&7&5&0&4&5&9&4&3&3&3\\
9&4&4&4&0&4&8&3&4&3&2\\
5&5&8&5&4&0&4&6&8&3&3\\
4&4&4&9&8&4&0&3&3&9&4\\
9&1&1&4&3&6&3&0&4&3&4\\
4&4&4&3&4&8&3&4&0&3&1\\
3&3&3&3&3&3&9&3&3&0&7\\
5&3&7&3&2&3&4&4&1&7&0
\end{bmatrix}$$

4 算法分析与结论

我们在 Windows 95 平台上用 C++ 编写了基于多维标度的挖掘关联规则算法和先验算法，并在模拟数据** 上对两者进行了测试. 因为先验算法要反复扫描数据库，迭代产生大量的后来证明不是数据项频集的数据项候选集，效率较低. 而基于多维标度的挖掘关联规则算法仅仅得到2位数据项频集即可，不需要再扫描数据库. 先验算法得到的是全部数据项频集，而多维算法只得到概要性的关联规则. 用户可以对其感兴趣的数据项再进行聚焦. 我们还进行了测试，结果表明，随着数据库规模的增大，多维标度算法具有很好的可扩展性.

---

* *apriori-gen*(L_1)与文献[1]中的 *apriori-gen*()算法相同.

** http://www.almaden.ibm.cm/cs/quest/syndata.html

参考文献

1 Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. In: Buneman P, Jajodia Sushil eds. Proceedings of the 1993 ACM SIGMOD International Conference Management of Data. Washington, DC: ACM Press, 1993. 207～216

2 Park J S, Chen M S, Yu P S. An effective hash-based algorithm for mining association rules. In: Windom J ed. Proceedings of the 1995 ACM SIGMOD International Conference Management of Data. San Jose, CA: ACM Press, 1995. 175～186

3 Houtsma M, Swami A. Set-oriented mining for association rules in relational databases. In: Yu lip S, Chen Arbee L P eds. Proceedings of the 11th International Conference'95 on Data Engineering. Taipei: IEEE Computer Society, 1995. 25～34

4 Klemttinen M, Mannila H *et al*. Finding interesting rules from large sets of discovered association rules. In: Adam N R eds. Proceedings of the 3rd International Conference on Information and Knowledge Management. Maryland: ACM Press, 1994. 401～407

5 Srikant R, Agrawal R. Mining generalized association rules. In: Umeshwar Dayal eds. Proceedings of the 21st International Conference on Very Large Data Bases. Zurich, Switzerland: Morgan Kaufmann Publishers, Inc., 1995. 407～419

6 Han J, Fu Y. Discovery of multiple-level association rules from large databases. In: Umeshwar Dayal eds. Proceedings of the 21st International Conference on Very Large Data Bases. Zürich, Switzerland: Morgan Kaufmann Publishers, Inc., 1995. 420～431

7 Mrikant R, Agrawal R. Mining quantitative association rules in large relational tables. In: Jagadish H V, Inderpal Singh Mumick eds. Proceedings of ACM SIGMOD International Conference'96 on Management of Data. Montreal, Canada: ACM Press, 1996. 1～12

8 Cheung D W, Han J, Ng V *et al*. Maintenance of discovered association rules in large databases: an incremental updating technique. In: Su Stanley Y W ed. Proceedings of the 12th International Conference on Data Engineering. New Orleans, Louisiana: IEEE Computer Society, 1996. 106～114

A Multidimensional Scaling Based Algorithm for Fast Mining Association Rules

XIAO Li　JIN Yuan-ping　XU Hong-bing　WANG Neng-bin

(*Department of Computer Science and Engineering　Southeast University　Nanjing*　210096)

Abstract　Mining association rules is a major aspect of data mining research. In this paper, on the basis of analyzing the basic properties of the problem of mining association rules and multidimensional scaling, the authors propose a new multidimensional scaling based algorithm for fast mining association rules, which projects each data item on multidimensional space according to the association scaling between data items, and then reduces the dimensions of the space. Finally, the algorithm generates association rules more focused and presents them in a visual style.

Key words　Data mining, knowledge discovery, association rules, multidimensional scaling technique, visualization.