doi: 10.3969/j.issn.1006-1576.2012.03.027

Apriori 改进算法在军队院校干部考核中的应用

许子君[1]，杜秋[1]，栾超[2]
（1. 军事经济学院研究生 4 队，武汉 430035；2. 中国人民解放军 69052 部队，乌鲁木齐 830002）

摘要：针对目前我军院校考评办法和考评体系存在的不足，对 Apriori 算法在军队院校干部考核中的应用进行研究。从数据挖掘算法的角度，通过实例分析 Apriori 算法在军队院校专业技术干部考评关联规则挖掘中的应用，并根据专业技术干部考评关联规则挖掘的特殊性提出 Apriori 关联规则算法在干部考评系统中应用的改进算法。结果表明：该方法能减少算法计算量，使挖掘更具针对性，可为院校的教学管理工作提供必要的决策支持。

关键词：数据挖掘；军队院校；信息管理；关联规则
中图分类号：TP311　　**文献标志码**：A

Application of Apriori Improved Algorithm in Military Academies Cadre Evaluation

Xu Zijun[1], Du Qiu[1], Luan Chao[2]
(1. *No*. 4 *Brigade of Graduate*, *Military Economic Academy*, *Wuhan* 430035, *China*;
2. *No*. 69052 *Unit of PLA*, *Urumqi* 830002, *China*)

Abstract: For the shortcomings of current military academies evaluation methods and evaluation system, research the application of Apriori algorithm for cadres assessment in the military academies. From the perspective of data mining algorithms, the author analyzed the application of Apriori algorithm in mining association rule evaluation of military academies professional and technical cadres by using instance, proposed improved Apriori in military academies cadre evaluation according to its specificity. The results show that this method can reduce the computational algorithms to make mining targeted, and can provide necessary management decision support for military institute teaching management.

Key words: data mining; military institute; information management; association rule

0　引言

院校专业技术干部考评系统，主要包括对专业技术干部进行年度和任期考核。自实行院校专业技术干部考评以来，对专业技术干部的教育、管理工作起到了良好的推进作用，同时也为干部部门提拔使用干部提供了科学的依据。但目前我军院校推行的专业技术干部职称考评还处于起步阶段，考评办法和考评体系还不是很完善，特别是经过几年的考评实践已经积累了大量的考评数据，但并没有对这些数据进行挖掘分析，从中找出有用的知识，从而造成大量有用信息的损失，不能为院校领导和干部部门提供决策支持。

为了从干部数据库和职称考评数据库中挖掘出专业技术干部工作绩效与干部属性的关联性，需选用关联规则算法。典型的关联规则挖掘算法是 R. Agralwal 等人提出的 Apriori 算法，它有效解决了传统关联规则算法中候选项目集大，计算量大等问题。现行的关联规则算法大多是以 Apriori 为核心，或是其变体，或是其扩展[1]。但因为 Apriori 算法在挖掘过程中需要多次扫描数据库，还是会产生大量的候选项目集，针对专业技术干部考评的特殊性，笔者结合院校现有专业技术干部管理数据库，对 Apriori 算法进行改进，并应用到专业技术干部考评中。

1　改进的 Apriori 关联规则算法

1.1　Apriori 关联规则算法

1.1.1　支持度和置信度的计算

在解释支持度和置信度的计算方法之前，必须要明确几个在关联分析中使用的基本术语。

关联规则挖掘的数据集记为 D（D 为事务数据库），$D=\{t_1,t_2,\cdots,t_k,\cdots,t_n\}$，$t_k=\{i_1,i_2,\cdots,i_j,\cdots,i_p\}$ $(k=1,2,\cdots,n)$为一条事务；t_k 中的元素 $i_j(j=1,2,\cdots,p)$ 称为项目（item）。设 $I=\{i_1,i_2,\cdots,i_m\}$是 D 中全体项目组成的集合，I 的任何子集 X 称为 D 中的项目集（item set），$|X|=k$ 称为 X 为 k–项集。设 t_k 和 X 分别为 D 中的事务和项目集，如果 $X\subseteq t_k$，称事务 t_k 包含项目集 X。例如，表 1 是购物篮的二元表示形式，其中每行表示对应一个事务，每列对应一个项。项

收稿日期：2011-09-20；**修回日期**：2011-10-24
作者简介：许子君（1984—），男，山东人，博士，从事军需勤务、后勤自动化研究。

可以用二元变量表示，如果项在事务中出现，则它的值为 1，否则为 0。每个事务 t_i 包含的项集都是 I 的子集，如果一个项集包含 k 个项，则称它为 k-项集。例如事务 1{面包，牛奶}是一个 2-项集。

表 1 购物篮的二元表示形式

TID	面包	牛奶	尿布	啤酒	鸡蛋	可乐
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

事务的宽度定义为事务中出现项的个数。如果项集 X 是事务 t_j 的子集，则称事务 t_j 包含项集 X，例如表 1 中第 2 个事务包括项集{面包，尿布}。项集的一个重要性质是它的支持度计数，即包含特定项集的事务个数。数学上，项集 X 的支持度计数 $\sigma(X)$ 可以表示为：

$$\Sigma(X)=|\{t_i|X\subseteq t_i,\ t_i\in T\}|$$

其中符号| |表示集合中元素的个数。项集{啤酒，尿布}的支持度计数为 3，因为有 3 个事务同时包含这 2 个项。

关联规则是形如 $X\rightarrow Y$ 的蕴涵表达式，其中 X 和 Y 是不相交的项集，即 $X\cap Y=\Phi$。关联关则的强度可以用它的支持度(support)和置信度(confidence)度量。支持度确定规则可以用于给定数据集的频繁程度，而置信度确定 Y 在包含 X 的事务中出现的频繁程度。支持度(support)和置信度(confidence)这 2 种度量的形式定义如下：

$$s(X\rightarrow Y)=\frac{\sigma(X\cup Y)}{|D|}\times 100\%$$

$$c(X\rightarrow Y)=\frac{\sigma(X\cup Y)}{\sigma(X)}\times 100\%$$

其中$|D|$是事务集 D 的事务数。若 support($X\rightarrow Y$)不小于用户指定的最小支持度(min support)，则称$\{X, Y\}$为频繁项目集，否则称$\{X, Y\}$为非频繁项目集。

支持度和置信度是描述关联规则的 2 个重要概念，前者用于衡量关联规则在整个数据集中的统计重要性，后者用于衡量关联规则的可信度。一般来说，只有支持度和置信度均很高的关联规则才可能是用户感兴趣、有用的关联规则。通常用户根据挖掘需要指定最小支持度(min support)和最小置信度(min confidence)。前者描述了关联规则的最低重要程度，后者规定了关联规则必须满足的最低可靠性[2]。

1.1.2 Apriori 关联规则算法

Agrawal 等首先提出了挖掘顾客交易数据库中项集间的关联规则问题，其核心方法是基于频集理论的推导方法。可将关联规则挖掘算法设计分解成 2 个子问题。

1) 找到所有支持度大于最小支持度的项集(item set)，这些项集就是前面探讨过的频繁项集。

2) 使用第 1 步找到的频集，产生期望的规则。

第 2 步相对简单一点。如给定了一个频集 $Y=i_1, i_2,\cdots,i_k$, $k>1$, $i, j\in i$，产生只包含集合 $\{i_1,i_2,\cdots,i_k\}$ 中项的所有规则(最多 k 条)。其中每一条规则的右部只有一项，即形如$[Y-i_i]\Rightarrow I_i$, $1\leq I\geq k$。一旦生成这些规则，则只保留这些大于用户给定最小置信度的规则。

关于 Apriori 算法可以具体描述如下：

输入：事务数据库 D；最小支持度 min support

输出：D 中的频繁项集 F。

算法：

```
1:  k=1
2:      FK= {i | i ∈ I ∩ σ({i}) ≥ min sup}
//发现所有的频繁 1-项集
3:  repeat
4:      k=k+1
5:      CK=apriori-gen(FK-1)//产生候选项集
6:      for 每个事务 t∈T do
7:          CT=subset(Ck,t)
//识别属于 t 的所有候选
8:          for 每个候选项集 c∈CT do
9:              σ(c)=σ(c)+1   //支持度计数增值
10:         end for
11:     end for
12:     FK={c|c∈CK ∩σ(c)≥min sup}
//提取频繁 k-项集
13:     until FK= ∅
14:     Result = ∪FK[3]
```

Apriori 算法是一种宽度优先算法，通过对数据

库 D 的多趟扫描来发现所有的频繁项目集。在每一趟 k 中只考虑具有同一长度 k(项目集中所含项目的个数)的所有项目集。在第一趟扫描中，Apriori 算法计算所有单个项目的支持度，生成所有长度为 1 的频繁项目集。在后续的每一趟 k 中，首先以前一趟中所发现的所有频繁项集为基础，生成所有新的候选项目集(candidate item sets)，即潜在的频繁项目集，然后扫描数据库 D，计算这些候选项目集的支持度，最后确定候选项目集中哪一些真正成为频繁项目集。重复上述过程，直到再也发现不了新的频繁项目集为止[4]。

Apriori 算法之所以比较成功，重要的是它采用了 Apriori-gen 过程，从而可产生相对较小的候选集。函数 Apriori-gen 的候选产生过程为：合并一对频繁(k-1)-项集，仅当它们的前 k-2 个项都相同且第 k-1 个项不相同时，从而产生长度为 k 的候选项目集。

Insert into C_k
　Select P.item1,P.item2, … P.item(k-1), Q.item(k-1)
　From $F_{k-1}P$, $F_{k-1}Q$
　Where P.item1=Q.item1 and … P.item(k-2) =Q.item(k-2) and P.item(k-1)<>Q.item(k-1)

其中 P，Q 为频繁(k-1)项集。由于每个候选都由一对频繁(k-1)项集合并而成，因此需要附加的候选剪枝步骤来确保该候选的其余 k-2 个子集是频繁的。因为一个频繁项目集的任何子集一定也是频繁项目集，所以采用如下算法对候选项目集进行剪枝。

For all $C \in C_K$ do
　For all(c 的包含 k-1 个项目的子集 s) do
　If(s 不属于 F_{k-1})then
　delete from C_K
　end if
　end for
end for

算法高效率的关键在于生成较小的候选项目集，也就是尽可能不生成和计算那些不可能成为频繁项集的候选项目集[5]，这样大大减少了数据挖掘中的计算量，提高了关联规则的挖掘效率。

1.2　改进的 Apriori 关联规则算法

对于海量的挖掘数据来说，Apriori 关联规则算法所带来的计算量仍然很惊人，特别是当事务和候选项集的数目非常大时。例如：假设算法得到的 1 项频繁集的数量是 10^4，则根据 Apriori 算法将会产生 10^7 个 2 项候选集，由于 2 项候选集没有剪枝，所有这些候选集都需要校验，由此带来的计算量是惊人的[6]。Apriori 算法在大量候选集产生的情况下基本很难运行。

将 Apriori 关联规则算法应用到院校专业技术干部考评系统中的目标是，找出专业技术干部工作绩效与干部的职称、学历、年龄、学缘结构、任职经历、立功情况等方面的关联。如果将 Apriori 算法简单地应用到这些关联规则的挖掘上面，不仅给挖掘工作带来了惊人的计算量，并且将会产生很多我们不想关注的关联规则，例如：

年龄 40 岁以上⇒教授
年龄 30 岁以下⇒讲师
学历结构博研 & 年龄 40 岁以上⇒教授

为了大大减少挖掘工作的计算量，并且能使挖掘工作更具目的性，需要对 Apriori 算法做些改进。关于 Apriori 算法在干部考评系统中应用的改进算法可描述如下：

输入：事务数据库 D，最小支持度 min support，最小置信度 min confidence
输出：频繁项集 F。
设考评成绩项目表示为：i_t

1：　k=1
2：　$F_K=\{i \mid i \in I \cap \sigma(\{i\}) \geq \min\sup\}$
//发现所有频繁 1-项集
3：　k=k+1
4：　$F_K=\{i,i_t \mid i \in F_1 \cap \sigma(\{i,i_t\}) \geq \min\sup\}$
//发现所有频繁 2-项集
5：　repeat
6：　k=k+1
7：　C_K=apriori-gen(F_{K-1})　//产生候选项集
8：　for 每个事务 t∈T do
9：　C_T=subset(C_k, t)
//识别属于 t 的所有候选
10：　for 每个候选项集 $c \in C_T$ do
11：　$\sigma(c)=\sigma(c)+1$　//支持度计数增值
12：　end for
13：　end for
14：　$F_K=\{c \mid c \in C_K \cap \sigma(c) \geq \min\sup\}$
//提取频繁 k-项集
15：　until $F_K=\varnothing$
16：　Result = $\cup F_K$

产生 k 项候选集的函数 apriori-gen 改进后如下：

```
Insert into Ck
    Select P.item1,P.item2, … P.item(k-2), Q.item(k-2),it
    From FK-1P,FK-1Q
    Where P.item1=Q.item1 and … P.item(k-3)=Q.item(k-3) and P.item(k-2)<>Q.item(k-2)
对候选项集 CK 的剪枝算法修改如下：
For all C∈CK do
  If(k>3)
    For all(c 中包含 it 项的 k-1 个项目的子集 s) do
      If(s 不属于 Fk-1)then
        delete from CK
      end if
    end for
  end if
end for
```

由于改进后的 Apriori 算法极大地减少了候选频繁项集的数量，只考虑笔者想要关注的频繁项集，故能大大减少关联规则挖掘的计算量。在候选集以及频繁项集的生成过程中，将考评成绩项 i_t 作为所有频繁项集和候选集(除 1-项集以外)的子集，并使用 Apriori 关联规则的改进算法对院校专业技术干部考评系统进行关联规则分析。

2 数据准备阶段

数据准备阶段所得到的挖掘数据的质量是数据挖掘成败的关键因素。数据准备阶段的工作主要包括：一是从多个数据源去综合数据挖掘所需要的数据集，并且要保证数据的综合性、易用性、数据的质量和数据的时效性。二是要清理数据，就是要发现并清理数据集中存在的错误、异常或缺失的数据。三是从现有的数据集衍生出适于关联规则挖掘的数据集，这就涉及到数据的转换工作。

笔者从某院校干部管理数据库(gbzh 数据库)和专业技术干部考评系统数据库(gbkp 数据库)中抽取所有专业技术干部基本情况数据及一个考评周期内考评成绩数据，导入新建表“A_技术考评”中。

数据转换的目的在于将关联分析原始数据集转换成适于关联规则挖掘的数据集。通过前面对 Apriori 关联规则算法的分析，可以得出用于关联规则挖掘的数据必须满足 2 个条件：数据是布尔型的，布尔型数据的维数不能太大。

1) 部别。部别对应的数据是布尔型，但是数据的维数太大，不适于关联规则挖掘，笔者将部别划分为 3 类：基础部、专业系和机关，分别标记为：B_1，B_2，B_3。代码为：

```
update A_技术考评
set 部别=
case
    when 部别 like '基础部%' then 'B1'
    when 部别 like '%系%' then 'B2'
    when 部别 not in('B1', 'B2') then 'B3'
end
```

2) 性别、文化程度、技职类别、考评成绩。之所以将这 4 项列在一起是因为这 4 项已经完全符合用于关联规则挖掘的数据必须满足的条件，只需要将其做相应标记即可：性别(男：S_1，女：S_2)；文化程度(博研：W_1，硕研：W_2，本科：W_3，大专：W_4，中专及以下：W_5)；技职类别(正高：J_1，副高：J_2，中职：J_3，低职：J_4)；考评成绩(优秀：K_1，称职：K_2，基本称职：K_3，不称职：K_4)。

3) 出生时间、工作时间。出生时间和工作时间属于连续变量，首先要对其离散化，然后做相应标记。笔者将年龄划分为 3 个层次(30 岁及以下，31～40 岁，41 岁及以上)，对应的出生日期就为(1978 年及以后，1968～1977 年，1967 年及以前)，分别标记为：N_1，N_2，N_3，并放入新建字符字段“年龄”里。对应代码如下：

```
update A_技术考评
set 年龄=
case
    when 出生时间>=to_date('1978-01-01', 'yyyy-mm-dd') then 'N1'
    when 出生时间<to_date('1978-01-01', 'yyyy-mm-dd') and 出生时间>=to_date('1968-01-01', 'yyyy-mm-dd') then 'N2'
    when 出生时间<to_date('1968-01-01', 'yyyy-mm-dd') then 'N3'
end
```

以同样的方法将工作时间划分为 3 个层次：10 年工作年龄、20 年工作年龄以及 30 年工作年龄，分别标记为：G_1，G_2，G_3。并放入新建字符字段“工作经历”里，这里不再赘述。

4) 毕业院校。毕业院校对应的数据虽然是布尔型的，但数据维数还是太大，也不适于关联规则的挖掘，笔者将毕业院校划分为 2 类：本校毕业和异校毕业。分别标记为：Y_1，Y_2，代码如下：

Update　　A_技术考评
Set　　毕业院校='Y_2'
Where　　毕业院校 not like ('某某学院%')

Update　　A_技术考评
Set　　毕业院校='Y_1'
Where　　毕业院校 like ('某某学院%')

3　关联规则的挖掘

笔者设定最小支持度(min support)为：10%，最小置信度(min confidence)为 40%。根据 Apriori 关联规则改进算法，首先计算频繁单项集，所得到的结果如表 2。

表 2　频繁单项集

单项集合	支持度/%	单项集合	支持度/%
B_1	18	J_1	10
B_2	56	J_2	26
B_3	26	J_3	48
S_1	72	J_4	16
S_2	28	Y_1	31
N_1	26	Y_2	69
N_2	37	G_1	27
N_3	37	G_2	37
W_1	10	G_3	36
W_2	39	K_1	35
W_3	49	K_2	37
		K_3	20

下一步，连接频繁单项集并依据最小支持度和最小置信度生成频繁 2-项集。前面笔者介绍了获取频繁 2-项集的方法是：$F_K=\{i,i_t \mid i\in F_1\cap\sigma(\{i,i_t\})\geq \min\sup\}$(其中 i_t 属于考评成绩项集)，计算后得到的频繁 2-项集如表 3。

表 3　频繁 2-项集

双项集合	支持度/%	置信度/%	双项集合	支持度/%	置信度/%
B_2 & K_1	22	40	W_3 & K_2	25	51
B_3 & K_2	12	46	J_3 & K_2	21	44
N_2 & K_2	16	43	Y_2 & K_1	29	43
N_3 & K_1	19	51	G_1 & K_3	11	41
W_2 & K_1	18	46	G_3 & K_1	16	44

根据 Apriori 关联规则改进算法中的 Apriori-gen 过程，需要连接频繁 2-项集生成长度为 3 的候选项目集，再根据对支持度和置信度的计算获得频繁 3-项集，连接频繁 2-项集得到的候选 3 项集有：$B_2\&N_3\&K_1$；$B_2\&W_2\&K_1$；$B_2\&Y_2\&K_1$；$B_2\&G_3\&K_1$；$N_3\&W_2\&K_1$；$N_3\&Y_2\&K_1$；$W_2\&Y_2\&K_1$；$N_3\&G_3\&K_1$；$W_2\&G_3\&K_1$；$Y_2\&G_3\&K_1$；$B_3\&N_2\&K_2$；$B_3\&W_3\&K_2$；$B_3\&J_3\&K_2$；$N_2\&W_3\&K_2$；$N_2\&J_3\&K_2$；$W_3\&J_3\&K_2$。对候选 3 项集进行支持度和置信度的计算得到频繁 3 项集如表 4。

表 4　频繁 3-项集

3 项集合	支持度/%	置信度/%	3 项集合	支持度/%	置信度/%
$B_2\&Y_2\&K_1$	18	50	$N_2\&J_3\&K_2$	12	41
$N_3\&G_3\&K_1$	15	47			

现在，通过频繁 3-项集已经无法生成长度为 4 的候选项集，所以关联规则挖掘到此结束。

4　解释与评估

1) 通过分析单项频繁项集可以发现，高学历专业技术干部人才在学院专业技术干部队伍中占主导，其中博士及以上学历的支持度约为 10%，硕士学位的支持度约为 39%，本科学历的支持度约为 49%，说明学院近年来坚持实施人才战略，重视高学位中青年专业技术干部的培养和引进已初具成效。

2) 频繁项集 $B_2\&K_1$ 表明专业系技术干部被评为优秀的概率相对其他部门要大。拿基础部做比较，该部门专业技术干部被评为优秀的置信度为：confidence($B_1\&K_1$)=0.6/0.18=33%，而专业系技术干部被评为优秀的置信度为：confidence($B_2\&K_1$)=40%。这与各专业系科研项目多、学术成果丰有较大的关系，另外院领导的重视方向也是其中的重要因素。

3) 频繁 2-项集 $N_3\&K_1$，$W_2\&K_1$，$G_3\&K_1$，以及频繁 3-项集 $N_3\&G_3\&K_1$ 表明年龄处在 40 岁以上、文化程度为硕士、工作经验丰富的专业技术干部科研水平较高，有宽广的知识面，能较好地把握学科的学术动态，能将教学和科研很好地结合起来，他们被评为优秀的概率比较大。

4) 频繁 3-项集 $N_2\&J_3\&K_2$，说明年龄处在 30～40 岁之间，技职类别为中职的被评为称职的概率比较大，说明学院一批中青年专业技术骨干已经成长起来，学院专业技术干部队伍结构也趋于合理化。

（下转第 96 页）