基于二进制关联规则提取算法的大学生就业竞争力分析

刘 斓 徐健锋
(南昌大学信息工程学院计算机系 南昌 330031)

摘 要 目前大学生就业竞争力分析是一个很重要的课题,但现有的分析方法多局限于对竞争力影响因素所占权重的分析,从而制约了大学生对各影响因素间的关联性的认识。根据关联规则分析理论,通过转化决策型数据,利用二进制关联规则挖掘算法对大学生就业竞争力进行科学分析,提取得出有效规则。通过对2007年度毕业生的就业竞争力评价,验证该方法的实用性、可行性。

关键词 数据挖掘,关联规则,BRT算法,就业竞争力

College Students' Employment Competition Power Analysis Based on Mining Algorithm of Binary Association Rules

LIU Lan XU Jian-feng
(Department of Computer, School of Informatin Engineering, Nanchang University, Nanchang 330031, China)

Abstract At present, many research of analysis systems of competition power methods for student obtaining employment is very importent competition ability question discussion. Currently, there are several these methods, but they are limited to the analysis of the weighting of the influenced fact weight aspects. Then the limitation restricts the cognition of college students to cognize scientificly the relation between these facts aspects. According to the analysis of association rules theory, we analyzed scientifically the competitive power of students by using the association rule mining algorithm in binary system. We may extract effective association rules through exmerimentation. Association rule mining analyzes college students' competiton power with traforming the decisiong-making statistics. The method was used to analysis of competitive power for graduations obtained students' employment this year, result is nice. It shows the methods are feasible and effective. The thesis validated Boolean rule, method's practicability and feasibility through the evaluating of the graduate students' employment competition ability this year.

Keywords Data mining, Aassociation rules, BRT algorithm, Obtaining employment, Competitive power

当前我国高等教育规模急剧扩大,高校毕业生就业已经越来越成为社会各界关注的热点问题。随着改革开放的不断深入,经济建设的不断加速,市场经济的不断完善,加之大学毕业生总量的增加,劳动力市场的不完善,使得大学生就业形势越来越严峻。面对严峻的就业竞争形势,为了解决大学生就业中的问题,需要客观、准确地评价自己的就业竞争力。关于这些问题,许多学者都进行了大量的调查研究[1],但这些研究大都是通过分析每条评价指标的权数来确定该指标在整个评价体系中的重要度,只是为大学生提供一些简单的就业指导。

本文将影响大学生就业竞争力的学生个体因素进行二进制数据转换,再利用二进制型的关联规则提取算法即BR-T算法进行分析挖掘,获得这些因素中存在的一些内在联系,使就业指导机构和大学生对就业竞争力有更加科学的认识,从而不断提高大学生的就业竞争力。

1 基本概念及算法

1.1 基本概念

设 $I=\{i1,i2,\cdots,im\}$ 是 m 个不同属性的集合,称为项集(Item set)[2];包含 k 个属性的项集称之为 k-项集。记 S 为事务的集合,事务 T 是项的集合,并且 T 为 I 的子集,即 $T\subseteq I$。对每一个事务有唯一的标识,记作TID。设 A 是 I 中一个项集,如果 $A\subseteq T$,那么称事务 T 包含 A。

定义1(决策规则) 形如 $A\rightarrow B$ 的蕴涵式,其中 $A\subset I$, $B\subset I$,且 $A\cap B=\emptyset$。

定义2(规则 $A\rightarrow B$ 的支持度 SUP($A\rightarrow B$)与支持数) 支持度是事务集合 S 中的项集 $A\cup B$ 在所有事务中出现的概率 $P(A\cup B)$,S 中满足 $A\cup B$ 的事务个数为支持数。

定义3(规则 $A\rightarrow B$ 的可信度 conf($A\rightarrow B$)) 可信度等于条件概率 $P(B|A)$。

定义4(频繁项集,Frequent Itemset) 是指满足最小支持度(min_sup)的项集。

定义5(关联规则) 同时满足最小支持度阈值(min_sup)和最小可信度阈值(min_conf)的规则,即 sup($A\rightarrow B$)>=min_sup 且conf($A\rightarrow B$)>=min_conf 成立时,规则 $A\rightarrow B$ 称之为关联规则。

1.2 基于二进制型的关联规则算法

定义6(属性的二进制数转换) 事务数据库中各个属性根据其在各个事务是否取值的情况转换二进值数,即如果某事务中该属性有取值,则该二进制位的数据记为1,否则记为

到稿日期:2008-05-30 本课题获中科院计算所开放课题[IIp2006-3],江西省教育厅科技资助项目(赣教技[GJJ080382008])资助。
刘 斓(1973—),女,讲师,硕士,研究方向为数据挖掘、软件工程,E-mail:liulanxf@sohu.com。

0。并根据各个事务的次序连成一个二进制数串，即完成了各个属性相应的二进制数据转化。而各属性构成的1项集的支持数可通过计算各个属性二进制数序列中的1的个数获得。

定义7 BR-T由一个标号为Root的根结点和数个树结点构成，每一个结点可带有 n 个子树结点($n=0,1,2,\cdots$)。当 $n=0$ 时，称该结点为叶结点。一个事务数据库由一棵BR-T树表示，节点到根Root的路径上所有结点组合代表不同属性的集合中的项集。这项集的支持数由这个集合中各个属性的二进制数进行按位与运算后计算1的个数获得。

(注：在D中事务数量一定的情况下，可以由支持数代表支持度。下文中都用此方法表示)

定义8 BR-T除根结点外，每一个结点由Ii. cell，Ii. count和Ii. pointer 3个基本的域构成。其中Ii. cell代表某属性的标号，Ii. count为该结点表示的item set支持数，li. pointer为指向其父结点的指针。

定义9(候选频繁模式集CF) 由频繁 k 项集组成的集合。

1.3 BR-T关联规则算法步骤

1)构造BR-T树

步骤1 根据定义6将事务数据库 S 的每个属性分别进行二进制数转化。

步骤2 通过计算每个属性转化的二进制数串中1的个数，即可获得各个属性的支持数。当支持数小于某事先约定的阀值时该属性可被约简。

步骤3

步骤3.1 首先创建BR-T树的根节点root。将事务数据库 S 的频繁一项集，根据支持数的大小将各个属性依次插入到BFP树中，作为树根的各个叶节点。

步骤3.2 各个叶节点表示的 k 项集的 k 个节点(属性)的二进制数，与其右兄弟各树叶属性二进制数分别进行二进制数与运算。计算后获得二进制数中1的个数即为 $k+1$ 项集的支持数。当支持数大于阀值时，新属性插入上述 k 项集的树叶后，成为下一层的新叶节点。新叶节点代表次加入新属性的 $K+1$ 项集。

步骤3.3 重复记录插入树的过程步骤3.2，直到各个子树的各个叶子都没有一项集插入到BR-T树中。由此，事务数据库S转换成BR-T树。

步骤4 建立一个队列TL，保存从左到右各个叶节点地址。

2)挖掘BR-T树过程

步骤1 取TL表中的首元素，得到该元素指向的叶结点表示的项集并送入频繁项集CF。同时提取该叶节点的父节点指针，如果TL没有相同指针记录，则父节点指针加入队列TL。然后队列TL首元素出队。

步骤2 再取表TL的首元素，重复上述步骤1直到TL中的元素都被取出。

步骤3 最后可根据定义3计算各候选关联规则的置信度，并且依照定义5用给出的最小置信度筛选出所需要的关联规则。

注：计算置信度的过程略。

3)算法简要分析

设事务数据库 S 的属性个数为 k，记录数量(即项集数量)为 n。则本算法的最坏时间复杂度为 $O(2^k-k)$，最优时间复杂度为 $O(k^2)$。可见BR-T树算法与项集个数无关，而与属性个数成指数级相关。

2 大学生就业竞争力关联规则挖掘

2.1 评价体系与数据获取

关系数据库一般可分为无决策型和决策型。无决策型关系数据库中所有属性的地位相同，而决策型关系数据库中的属性则分为条件属性和决策属性[3]。本文以大学生初次就业竞争力作为研究对象，调查中由于要得出大学生竞争力的强弱，因此竞争力即为决策属性，竞争力强记为D，竞争力弱记为D′。

综合测评的要素即条件属性包括：

a：社会实践能力强；

a′：社会实践能力弱；

b：学习能力强；

b′：学习能力弱；

c：诚实守信等非智力的因素强；

c′：诚实守信等非智力的因素弱；

d：社会关系强；

d′：社会关系弱。

利用上述设计的调查表，我们在南昌大学软件工程专业07界毕业生中进行了调查，获得实验原始数据。整理后获得决策型事务数据集，如表1所列。

表1 大学生就业竞争力事务数据集S

TID	a	a′	b	b′	c	c′	d	d′	D	D′
1	1	0	0	1	1	0	1	0	1	0
3	0	1	1	0	1	0	1	0	1	0
5	1	0	1	0	0	1	0	1	1	0
7	1	0	1	0	0	1	1	0	1	0
9	0	1	1	0	0	1	0	1	0	1

1)构造BR-T树

设定支持度阀值为0.3，即最小支持数为3。

步骤1 根据定义6将表1事务数据集 S 的每个属性分别进行二进制数转化。

步骤2 获得各个属性的支持数。当支持数为3时该属性可被约简。

整理后各个属性的二进制数以及相应的1项集的支持数如表2所列。

表2 各属性的二进制数及支持数

属性	二进制数	支持数
D	1110111100	7
b	0111101010	6
d′	0101110011	6
b′	1000010101	4
d	1010001100	4

步骤 3

步骤 3.1　首先创建 BR-T 树的根节点 Root。将事务数据集 S 的频繁一项集，根据支持数的大小将各个属性依次插入到 BR-T 树中，作为树根的各个叶节点。结果如图 1 所示，用孩子兄弟法表示的第一层 BR-T 树由根节点 root 与它的 10 个孩子组成。

注：以下各图皆用孩子兄弟表示法表示 BR-T 树。

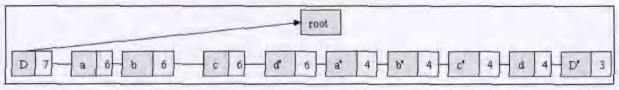

图 1　用孩子兄弟化表示的第一层 BR-T 图

步骤 3.2　各个叶节点表示的 k 项集的 k 个节点（属性）的二进制数，与其右兄弟各树叶属性二进制数分别进行二进制数与运算。计算后获得二进制数串中 1 的个数即为 $k+1$ 项集的支持数。当支持数大于阀值时，新属性插入上述 k 项集的树叶后，成为下一层的新叶节点。新叶节点代表次加入新属性的 $K+1$ 项集。例如图 1 中叶节点 D 分别与其右兄弟进行二进制数与运算获得 $(D \cup a)$ 的支持数为 6，大于最小支持数 3。所以在第一层节点 D 下面插入新节点，加入 a 以及 $(D \cup a)$ 的支持数 6 的信息。节点 D 分别与其他右兄弟 $\{b, c, d', a', b', c', d, D'\}$ 进行上述处理后的 *BR-T* 树如图 2 所示。

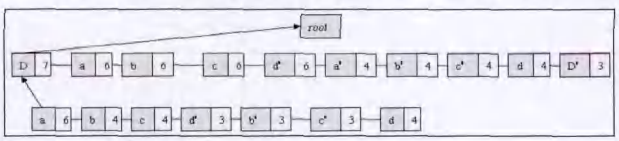

图 2　建树过程中的 BR-T 树

步骤 3.3　各子树的叶节点重复记录插入树的过程步骤 3.2，直到各个子树的各个叶子都没有一项集插入到 BR-T 树中。由此，大学生就业竞争力事务数据集转换成图 3 所示的 BR-T 树。

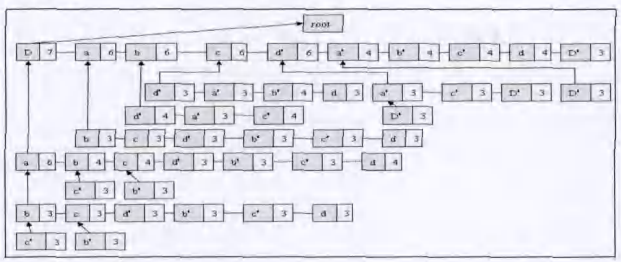

图 3　建 BR-T 树完成

注：由于本数据集是决策型数据集，某子树中没有决策属性 D 或 D' 的节点则该子树各个叶节点不需要进行步骤 3.2。例如图 3 所示的 BR-T 中以第一层节点 a, b, c 为根节点的子树，由于没有任何节点是属性 D 或 D'，因此可以不继续步骤 3.2 即可剪枝。

步骤 4　建立一个队列 TL 保存从左到右各个叶节点地址。

2）挖掘 BR-T 树过程

步骤 1　在队列 TL 中执行出队的操作，得到该元素指向的叶结点表示的项集和支持数，如果该项集中有属性 D 或 D' 则送入频繁项集 CF。同时提取该叶节点的父节点指针，如果 TL 没有相同指针记录，则父节点指针入队列 TL。

步骤 2　重复上述步骤 1 直到 TL 为空。获得的频繁项集如表 3 所列。

步骤 3　从表 3 中可转化为 10 条候选关联规制如下表 4 所列。本例设定最小置信度为 1，则其中序号 2，3，4 三条关联规制被过滤，最后可获得 7 条关联规制。

表 3　频繁项集 CF

频繁集	支持数
(d′,a′,D′)	3
(b,c′,D)	3
(c,b′,D)	3
(a,b,D)	3
(a,c,D)	3
(a,d,D)	3
(a,b′,D)	3
(a,c′,D)	3
(a,d′,D)	3
(a,b,c′,D)	3
(a,c,b′,D)	3

表 4　获得的关联规则

序号	候选关联规制	置信度	支持数
1	(d′,a′)→D′	1	3
2	~~(b,c′)→D~~	~~3/4~~	~~3~~
3	~~(c,b′)→D~~	~~3/4~~	~~3~~
4	~~(b,c′)→D~~	~~3/4~~	~~3~~
5	(a,c)→D	1	3
6	(a,d)→D	1	3
7	(a,b′)→D	1	3
8	(a,c′)→D	1	3
9	(a,d′)→D	1	3
10	(a,b,c′)→D	1	3
11	(a,c,b′)→D	1	3

通过与学生实际情况的验证，以上获得的关联规制符合客观情况。例如最后结论体现了社会实践能力强的学生如果还具备一项其他优势，则就业能力乐观。社会实践能力较弱的学生，如果又不具备较好的社会关系则，其就业前景堪忧。本结果对本学院大学生的就业能力培养具有较科学的参考价值。

结束语　Agrawal 等于 1993 年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，设计了基于频繁集理论的 Apriori 算法。其核心思想是基于频繁集理论的递推方法[2]。Apriori 算法在计算有效性上是公认最好的，但该算法可能存在产生大量的候选集，以及需要重复扫描数据库的两大缺点。Jiawei Han 等人提出的 FP_ growth 算法[5]的关联规则挖掘效率比 Apriori 算法高，但是它仍然需要扫描二次事务数据库。第一次扫描事务数据库，得到 L 表；第二次扫描事务数据库，构造出 FP-tree。但在实践应用中，当项集数量较大时，二次扫描实际的事务数据库的开销很大，因此减少数据库扫描次数能更好提高效率。基于上述研究，我们提出的 BR-T 算法有效地改善了上述问题，同时本文利用关联规则在大学生就业竞争力调查数据中进行挖掘，得出了一些有用规则，这些规则对判断大学生的就业竞争力的评价具有显著效果。另外，可以将决策属性置换成其他事务，例如工作薪酬等，以期得到更多对大学生就业有帮助的规则。同时如何与其他智能算法相结合也是下一步的研究工作。

参 考 文 献

[1] 李冬红，毛静，朱凌云. 大学生就业竞争力的模糊综合评判[J]. 中国大学生就业，2005(2)：61-62

[2] 陈文伟，黄金才. 数据仓库与数据挖掘[M]. 北京：人民邮电出版社，2004：143-151

[3] 白秀玲，崔林，王向阳，等. 关系数据库中关联规则的挖掘[J]. 电脑开发与应用，2002，15(10)：5-6

[4] Agrawal R，Srikant R. Fast Algorithms for Mining Association Rules[C]// Proc. of the-20st-VLDB Conference. Santiago，Chile，1994：487-499

[5] Han J，Pei J，Yin Y. Mining frequent patterns without candidate generation[C]// Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD' 00). 2000(5)：1-12