

一种基于生成树的分类去除关联规则冗余方法

楼巍¹, 邓柳燕¹, 严利民², 郭丽媛²

(1. 上海大学 机电工程与自动化学院, 上海 200072;

2. 上海大学 微电子研究与开发中心, 上海 200072)

摘要:在信息及数据爆炸的时代,冗余问题已经成为数据挖掘者获得知识的重大障碍,而目前解决的方法会导致关联规则的不完整性。基于此,文中引入了有向超图表示关联规则,重定义了邻接矩阵,介绍了冗余规则分类处理思想,将冗余规则分为从属规则和重复路径规则,通过VB编程去除了从属规则冗余,以及利用生成树算法去除了重复路径规则冗余。实验结果证明,此方法创新性地结合了图论中有向超图、生成树与关联规则的知识,维护了关联规则的完整性和准确性,同时去除了全部冗余规则。

关键词:关联规则;有向超图;邻接矩阵;生成树;冗余;去除从属规则

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2014)06-0024-04

doi:10.3969/j.issn.1673-629X.2014.06.006

A Method of Removing Redundant Association Rules by Classification Based on Spanning Tree

LOU Wei¹, DENG Liu-yan¹, YAN Li-min², GUO Li-yuan²

(1. School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200072, China;

2. Microelectronics R&D Center, Shanghai University, Shanghai 200072, China)

Abstract: In the information and data explosion era, redundant rules problem has become a major obstacle to gain knowledge for data miners, but the current solution may lead to the incompleteness of association rules. For the issues above, introduce the directed hypergraph to express association rules, redefine the adjacency matrix, propose the categorical thinking for removing redundant rules, dividing redundant rules into slave rules and repeated-path-rules, the algorithm of removing the dependency rule by VB programming and the spanning tree algorithm to remove the redundancy of repeated-path-rules. The result of experiments show that this method is effective, fast and to maintain the integrity and accuracy of association rules which links directed hypergraph, spanning tree of graph theory and association rules innovatively. At the same time it removes all redundant rules.

Key words: association rules; directed hypergraph; adjacency matrix; spanning tree; redundancy; removing slave rules

0 引言

随着数据量的激增以及数据复杂性的提升,不仅提高了在得到知识上的难度,更给获取有用知识提出了前所未有的考验。因为在数据激增情况下,通常产生的冗余规则数量远远大于有意义的规则的数量,这样规则的冗余问题便成为影响用户分析和有效利用关联规则的瓶颈^[1],而其研究价值在现今的信息时代不言而喻。

目前解决的方法^[2-3]主要有基于约束条件的挖掘、删除冗余规则、聚类、概括等,但这些方法是通过约

束挖掘条件来减少冗余规则,极有可能会漏掉一些有用规则,导致了关联规则的不完整性或者是获取信息的不完整性。

基于此文中介绍了一种基于生成树的分类去除关联规则冗余方法。这是一种全新的关联规则冗余去除法,它提出在有向超图的基础上,首次实现冗余分类处理,编程得到去除从属规则算法并结合了已有生成树算法,在不减少关联规则量的情况下,快速地去除了冗余规则,维护了关联规则的完整性和准确性。而其中有向超图因其便于表达多个选择子之间的依赖关系而

成为研究冗余规则问题的有力工具,生成树也因其连通性和无环路正好达到去除冗余的效果。

此方法在关联规则去冗余这一领域上首次提出分类去除冗余规则思想,并创新性地结合了有向超图和生成树。

1 冗余规则及邻接矩阵重定义

1.1 冗余规则

冗余规则^[4]一般可分为二种形式:
其一为从属规则,即规则 X_i 与 X_j 结论相同,而 X_i 的前提是 X_j 前提的充分条件或者反之,则 X_j 为冗余,重复规则可视为从属规则的特殊情况。
其二为重复路径规则,如果在规则库中存在选择子 X_i 、 X_j ,且 X_i 与 X_j 之间存在至少两条路径,则可判定存在冗余规则。

从属规则可由下列规则表示:
 $X_2 \rightarrow X_4$ (1)
 $X_2X_3 \rightarrow X_4$ (2)
 $X_3 \rightarrow X_4$ (3)
 $X_3 \rightarrow X_4X_5$ (4)
由规则(1)和(2)可以看出,两条规则的后项相同,前项存在交集,便认为规则(2)是冗余规则,则删除规则(2),保留规则(1)即保留前项中选择子少的一方,其中规则(1)和(2)两条成为从属规则。
由规则(3)和(4)可以看出,两条规则的前项相同,后项存在交集,便认为规则(4)是冗余规则,则删除规则(4),保留规则(3)即保留前项中选择子少的一方,其中规则(3)和(4)两条成为从属规则。

重复路径规则可由规则(5)和(6)表示:
 $X_1 \rightarrow X_2X_3 \rightarrow X_4$ (5)
 $X_1 \rightarrow X_5 \rightarrow X_4$ (6)
由规则(5)和(6)可知, X_1 到达 X_4 有两条路径,就认为路径重复,删除其中一条。

文中用有向超图^[5]表示关联规则,有向超图因为边的顶点可包含多个选择子,同时加上方向性可形象准确地表示关联规则,在此基础上去除冗余规则后也可简单还原。

定义1:有向超图是 $\vec{H} = (V, E)$ 二元对, V 是顶点集, E 是有向超边集。有向超边 $e \in E$ 被定义为有序对 $(T(e), H(e))$, 其中 $T(e), H(e)$ 均为 V 的子集, 且 $T(e) \cap H(e) = \emptyset, H(e)$ 称为有向超边的尾节点, $T(e)$ 称为有向超边的头节点。
用有向超图的头节点表示规则的前项,有向超图的尾节点表示规则的后项,其中规则的前项和后项可由多个选择子构成^[6-7]。

1.2 邻接矩阵重定义

在介绍文中分类去除关联规则冗余方法之前,必须在以往图论中的邻接矩阵概念上,再重新定义邻接矩阵^[8]。因图论中邻接矩阵主要应用于简单图中,而文中使用的有向超图较简单图的优势在于存在复合点,这使得复合规则只能用有向超图表示。文中定义的邻接矩阵是将复合点和简单点一样作为一行或者一列,这样能准确表达关联规则的信息,同时确保邻接矩阵为0-1矩阵以消除非0-1矩阵带来的算法处理难度以及表达的歧义,并且通过算法去除冗余后得到的邻接矩阵可简单还原成关联规则。文中定义的邻接矩阵及其对应有向超图如图1所示(其中图1的下图中虚线表示其两端的两个选择子结合为关联规则的前项,箭头指向为关联规则的后项)。

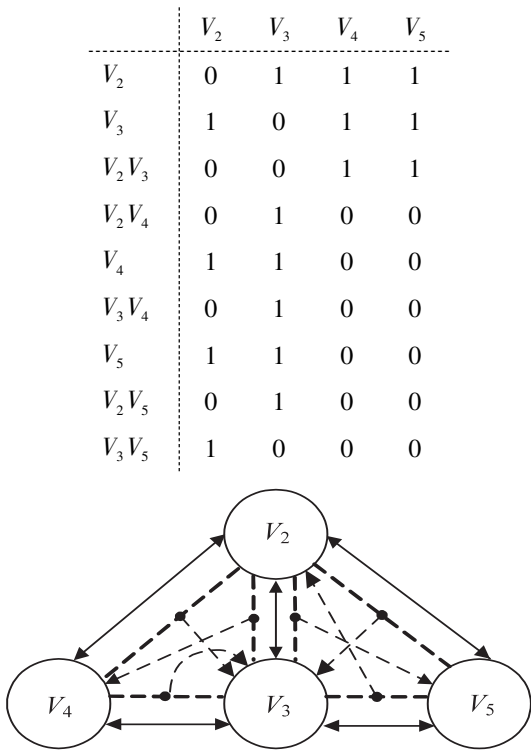


图1 重定义邻接矩阵及其对应有向超图

2 具体算法介绍

2.1 算法流程

文中采用的一种基于生成树的分类去除关联规则冗余方法的流程图如图2所示,具体步骤如下:
1)对实验数据进行扫描并生成关联规则,用有向超图表示关联规则,重新定义并得到其邻接矩阵。
2)重定义的邻接矩阵通过去除从属规则算法得到预处理的邻接矩阵。
3)得到预处理的邻接矩阵后通过生成树算法得到无环路且连通的生成树。
4)将得到的生成树的邻接矩阵还原成对应的关

联规则,得到最后的处理结果。

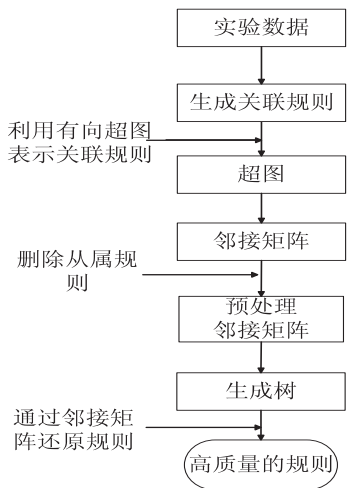


图 2 关联规则冗余检测流程图

2.2 去除从属规则算法

文中的去除从属规则算法把每一条关联规则定义成有向超图的一条边,根据上节得到重定义邻接矩阵,邻接矩阵的行表示关联规则的前项,邻接矩阵的列表示关联规则的后项^[9]。其算法流程图如图 3 所示,算法如下:

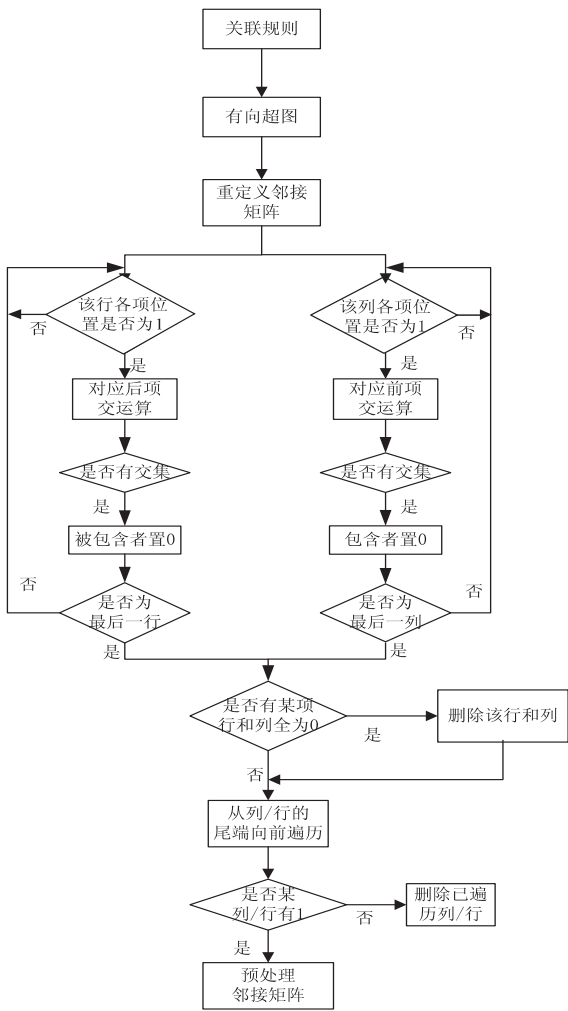


图 3 去除从属规则流程图

1)对于邻接矩阵每一列(关联规则后项),查询是否为 1,为 1 的位置对应的行(关联规则前项)进行交运算,若存在交集,则将对应行(关联规则前项)中包含项较多者为 1 的位置置 0,同理,对于每行处理如上^[10]。

2)遍历是否某一项处于的行和列值全为 0,若是,则删除该行和列。

3)最后,对于列和行分别从尾端遍历,若某列/行全为 0,删除该列/行,直到遇见存在 1 的列/行。

4)输出的矩阵即删除从属规则后的邻接矩阵,该邻接矩阵中的复合节点因为和其他节点无交集性,这就使得它在后续处理中独立性同简单节点一样,故可视为简单节点处理。

这样便删除了冗余规则里所有从属规则,得到预处理邻接矩阵。

2.3 生成树算法

去除从属规则后,冗余规则中还包含重复路径规则,为了去除这种冗余规则,文中引入了文献[11]中的 Kruskal 生成树算法。

在一个具有几个顶点的连通图 G 中,如果存在子图 G' 包含 G 中所有顶点和一部分边,且不形成回路,则称 G' 为图 G 的生成树

由生成树的概念^[11]可知,生成树具有连通且无回路的性质,那么去除从属规则算法后得到预处理邻接矩阵,然后通过生成树算法得到生成树,这时便已经去除关联规则中的重复路径规则。

文中虽引入了 Kruskal 生成树的算法,但因为数据挖掘目的是挖掘出有用的、潜在的知识,有些关联规则虽然支持度和置信度低(可用于作为最小生成树的权重矩阵),但是可能是很重要的知识,因此,文中算法去除了文献[10]中 Kruskal 算法的权重思想。因为生成树不是唯一的,不加权重的 Kruskal 算法可能会得到几种结果,而这几种结果都要作为等价考虑。Kruskal 算法思想是:

- 1)每个顶点各为一个子集。
- 2)每当选出一条边 (v, w) 的时候,判断它们是否已经在同一集合。
- 3)若在同一集合,那么就舍弃这条边。
- 4)若不在一个集合,选中这条边,将该边放入生成树边集中。同时合并 v 和 w 所在的顶点子集。

伪代码如下:

```
void Kruskal()
{
    int e=0; //记录边数
    每个顶点自成一个集合,并指定集合名。
    while(s<n-1) //n 为节点数
```

```
{
从 G 中选出当前最短边(v,w);
找到 v 所在集合的集合名 i,找到 w 所在集合的集合名 j;
if(i!=j)
{
将(v,w)加入树边集;
e++;
合并 i 和 j 集合;
}
}
```

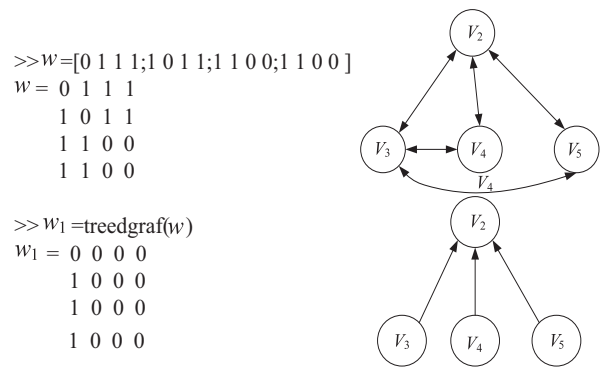


图 5 Matlab 仿真及其有向超图

表 1 去除冗余规则的结果

	Balloons	Shuttle-landing-control
关联规则总数	18	15
从属规则	8	9
重复路径规则	7	2
去除冗余规则数	15	11
剩余关联规则数	3	4

3 实验仿真结果

文中基于生成树的分类去除关联规则冗余方法主要有三个模块,分别是重定义邻接矩阵模块、去除从属规则模块和生成树模块,其中前两个模块通过 VB 编程实现,生成树模块在 Matlab 中实现,具体包含内容如图 4 所示。

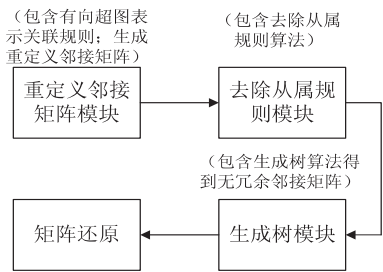


图 4 模块组成

为了验证该算法是否有效,文中选用了 2 个 UCI 数据集作为实验数据库,分别为 Balloons 数据集和 Shuttle-landing-control 数据集,其中 Balloons 数据集的最小支持度和最小置信度都设为 5%。

该数据集共有 4 个属性,通过 Aprior 算法^[12]共得到 18 条关联规则。

Shuttle-landing-control 数据集的最小支持度设为 40%,最小置信度为 100%。

该数据集共有 7 个属性,通过 Aprior 算法共得到 15 条关联规则。

以 Balloons 数据集为例,根据 18 条关联规则得到的重定义邻接矩阵,如图 1 所示。运行去除从属规则算法和生成树算法后得到无环路、连通的生成树,生成树算法具体过程及对应的有向超图如图 5 所示,多次运行生成树算法可能会有不一样的结果,每次结果都有参考价值。图 5 中左半部分是在 Matlab 中生成树算法的截图,可看出由预处理的邻接矩阵(去除从属规则算法结果)得到已经完全去除冗余的邻接矩阵,右半部分是其有向超图在生成树算法前后的变化。

最后,通过文中介绍的方法,两个数据集都准确且快速地去除了冗余规则,具体结果见表 1。

4 结束语

数据量的剧增使得关联规则挖掘产生大量的冗余规则,文中介绍了一种基于生成树的分类去除关联规则冗余的新算法,使用户能在不需要约束关联规则挖掘条件的情况下去除全部冗余。算法重新定义了邻接矩阵,突破性的将关联规则、有向超图、生成树算法结合起来实现分类去除冗余关联规则的思想。此算法在大数据时代也具有一定的优势,并且直接针对冗余,将其去除干净。实验仿真证明了这种新思路和方法是有效的、完整的且快速的。

参考文献:

[1] 马廷淮,张海盛,曾振柄.带结论域的关联规则的挖掘[J].计算机工程,2003,29(5):16-17.

[2] 张笑达,徐立臻.一种改进的基于矩阵的频繁项集挖掘算法[J].计算机技术与发展,2010,20(4):93-96.

[3] Shaw G,Xu Yue,Geva S.Utilizing non-redundant association rules from multi-level datasets[C]//Proceedings of 2008 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology. Washington,DC,USA:IEEE Computer Society,2008:681-684.

[4] Zaki M J.Mining non-redundant association rules[J].Data Mining and Knowledge Discovery,2004,9(3):223-248.

[5] 孙伟,郭莉,高天一,等.一种基于有向超图的规则库冗余及环路检测方法[J].大连理工大学学报,2008,48(1):74-78.

[6] Gallo G,Longo G,Palltino S,et al. Directed hypergraph and applications[J]. Discrete Applied Mathematics,1993,42(2-

股价误差都在0到2之间,体现该方法具有较强的预测能力。

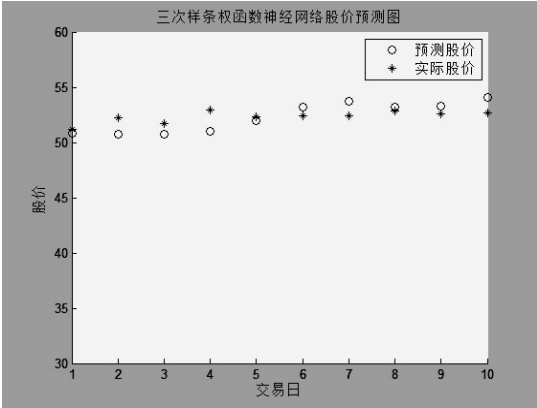


图5 三次样条权函数神经网络股价预测情况

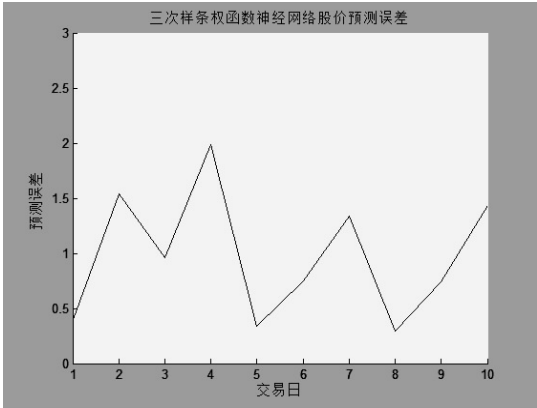


图6 三次样条权函数神经网络股价预测误差图

6 结束语

文中采用的三次样条权函数神经网络结构简单,只有两层,不含隐层,文中算法所需的神经元个数与样本数无关,克服了传统算法的诸多问题,网络的权值是与输入样本有关的函数,而不是常数。

基于样条权函数神经网络,文中提出了一种股票预测的新方法。所建立的模型简单,避免传统BP算法在网络训练时对隐层神经元个数、训练函数、激励函数等参数做出复杂的假设。实验表明基于此模型的预测方法预测能力良好,对于股市投资者具有一定的参

考价值。

参考文献:

[1] 沈冰. 股票投资分析[M]. 重庆:重庆出版社,2002.

[2] Schekman J A, Lebaran B. Nonlinear dynamics and stock returns[J]. Journal of Business, 1989, 62(3): 311-317.

[3] 陈之大, 贺学会. 证券投资技术分析[M]. 成都:西南财经大学出版社,1996.

[4] Guo Baolong, Guo Lei. A new approach to visual motion computation[J]. Journal of Xidian University, 1994, 21(4): 457-463.

[5] Saad E W, Prokhorov D V, Wunsch D C. Comparative study of stock trend prediction using time delay, recurrent and probability neural networks[J]. IEEE Trans on Neural Networks, 1998, 9(6): 1456-1470.

[6] Kohzadi N, Boyd M S, Kermanshahi B, et al. A comparison of artificial neural networks and time series models for forecasting commodity price[J]. Neurocomputing, 1996, 10(2): 169-181.

[7] Kuan Chung-Ming, White H. Artificial neural networks: an econometric perspective[J]. Econometric Reviews, 1994, 13(1): 1-91.

[8] Chan L W, Fallside F. An adaptive training algorithm for back propagation network[J]. Computers Speech and Language, 1987, 2(3-4): 205-218.

[9] Vogl T P, Mangis J K, Rigler A K, et al. Accelerating the convergence of the back propagation method[J]. Biological Cybernetics, 1988, 59(4-5): 257-263.

[10] Hsin H C, Li C C, Sun M, et al. An adaptive training algorithm for back propagation neural networks[C]//Proc of IEEE international conference on system, man and cybernetics. Chicago, IL: IEEE, 1992: 1049-1052.

[11] 张代远. 样条权函数神经网络的一种新型算法[J]. 系统工程与电子技术, 2006, 28(9): 1434-1436.

[12] 张代远. 新神经网络新理论与方法[M]. 北京:清华大学出版社, 2006.

[13] 林俊国. 证券投资学[M]. 北京:经济科学出版社, 2006.

[14] 赵宝福. 证券投资学[M]. 北京:中国物资出版社, 2000.

[15] 吴晓求. 证券投资学[M]. 北京:中国金融出版社, 2000.

(上接第27页)

3): 177-201.

[7] 崔阳, 杨炳儒. 超图在数据挖掘领域中的几个应用[J]. 计算机科学, 2010, 37(6): 220-222.

[8] 王志平, 王众托. 超网络理论及其应用[M]. 北京:科学出版社, 2008.

[9] Ramaswamy M, Sarkar S, Chen Ye-sho. Using directed hyper graphs to verify rule-based expert systems[J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9

(2): 221-237.

[10] Gursaran G S, Kaungo S, Sinha A K. Rule base content verification using a diagraph-based modeling approach[J]. Artif Intell Eng, 1999, 13: 321-336.

[11] 王海英, 黄强. 图论算法及其MATLAB实现[M]. 北京:北京航空航天大学出版社, 2010.

[12] 熊巧. Apriori算法的改进与应用[J]. 工业控制计算机, 2013, 26(4): 48-49.

一种基于生成树的分类去除关联规则冗余方法

作者:

[楼巍](#), [邓柳燕](#), [严利民](#), [郭丽媛](#), [LOU Wei](#), [DENG Liu-yan](#), [YAN Li-min](#), [GUO Li-yuan](#)

作者单位:

[楼巍, 邓柳燕, LOU Wei, DENG Liu-yan\(上海大学 机电工程与自动化学院, 上海, 200072\), 严利民, 郭丽媛, YAN Li-min, GUO Li-yuan\(上海大学 微电子研究与开发中心, 上海, 200072\)](#)

刊名:

[计算机技术与发展](#)

英文刊名:

[Computer Technology and Development](#)

年, 卷(期):

2014(6)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201406006.aspx