利用关联规则建立专家系统的知识库

张文静[1,3],宋 雨[1],卢海霞[2,3]

(1.华北电力大学 计算机科学与技术学院,河北 保定 071003;
2.华北电力大学 自动化系,河北 保定 071003;
3.河北农业大学 信息科学与技术学院,河北 保定 071001)

摘 要:数据挖掘和专家系统同属人工智能领域。关联规则是数据挖掘的一种方法,它的最典型的应用是超市的购物篮分析。专家系统主要解决的是智能推理问题而关联规则侧重于各个数据项之间有价值的联系。通过对关联规则的 Apriori 算法及规则的产生方法进行改动,挖掘出可应用于专家系统的知识库中的决策规则,从而找出了利用关联规则挖掘出用于决策的规则的方法。

关键词:数据挖掘;关联规则;Apriori 算法;专家系统;知识库
中图分类号:TP182 **文献标识码**:A **文章编号**:1005-3751(2005)02-0076-02

Making Knowledge Base of Expert System Through Application of Association Rules

ZHANG Wen-jing[1,3], SONG Yu[1], LU Hai-xia[2,3]

(1. School of Computer Science and Technology, North China Electric Power University, Baoding 071003, China;
2. Automation Department, North China Electric Power University, Baoding 071003, China;
3. College of Information Science & Technology, Agricultural University of Hebei, Baoding 071001, China)

Abstract: Both data mining and the expert system belong to the same artificial intelligence fields. Association rule is a method of data mining, whose typical application is the analysis of shopping basket in supermarket. The main task of expert system is to solve the problem of decision making, while association rule is to find out the valuable association of each items. By modifying the Apriori arithmetic and the method of the making rules, mine the decisive rule of database that could be applied to expert system, thereby find out the method of mining decisive rule using association rules.

Key words: data mining; association rule; Apriori; expert system; knowledge base

0 引 言

专家系统(Expert System)技术是以计算机为工具,利用专家知识及知识推理来理解与求解问题的知识系统[1]。

数据挖掘(Data Mining)是指从大型数据库或数据仓库中提取隐含的、未知的及有潜在应用价值的信息或模式[2]。关联规则是当前数据挖掘研究的主要模式之一,侧重于确定数据中不同领域之间的联系。

1 关联规则简介

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[3],包含以下两个主要步骤:

(1) 挖掘频繁项集:Apriori 算法[4]。

Apriori 算法利用了一个层次顺序搜索的循环方法来完成频繁项集的挖掘工作。首先利用 k - 项集来产生(k + 1) - 项集,然后经过 Apriori 算法的性质"一个频繁项集中任意一个子集也应是频繁项集"以及最小支持度的筛选,产生频繁 k - 项集,直到不能再产生频繁项集算法为止。

(2) 由频繁项集产生规则[5]:

具体产生关联规则的操作说明如下:

① 对于每个频繁项集 l,产生 l 的所有非空子集。

② 对于每个 l 的非空子集 s,若它的信任度不小于最小信任度阈值,则产生一个关联规则"$s \rightarrow (l - s)$"。

这是经典的关联规则的挖掘算法,下面通过举例来说明对该方法的改动,以及利用挖掘关联规则的方法来建立专家系统的知识库。

2 利用关联规则建立专家系统的知识库

在这里,对算法的改动主要体现在以下 3 个方面:首

收稿日期:2005-05-19

作者简介:张文静(1981—),女,河北邯郸人,硕士研究生,研究方向为软件工程、数据挖掘;宋 雨,教授,硕士生导师,研究方向为软件工程、软件构件/构架技术。

先是对挖掘数据项的建立方法的改动;第二是对算法连接规则的改动;第三是对产生规则的方法的改动。

下面以利用关联规则建立农业专家系统的知识库为例,分别说明以上的3种改动:

如,在农业专家系统中对于某一农作物病害的判断,它的几个病态的因素有:病斑颜色(黑褐色,粉红色和褐色),病斑位置(叶子病害,岭壳病害),病斑形状(圆形,半圆形,不规则),病斑特征(无特征,稍下陷特征和下陷特征)。根据这几个因素,最后得出病害的名称(炭疽病,印度炭疽病和角斑病)。

现根据以往经验得到如下一组数据:

1)黑褐色病斑,叶子病害,圆形,无特征→炭疽病

2)黑褐色病斑,叶子病害,圆形,无特征→炭疽病

3)粉红色病斑,岭壳病害,半圆形,稍下陷特征→炭疽病

4)褐色病斑,叶子病害,圆形,无特征→印度炭疽病

5)褐色病斑,叶子病害,圆形,无特征→印度炭疽病

6)褐色病斑,岭壳病害,圆形,下陷特征→角斑病

7)黑褐色病斑,岭壳病害,不规则,无特征→角斑病

8)黑褐色病斑,叶子病害,圆形 ,无特征→炭疽病

9)粉红色病斑,岭壳病害,半圆形,稍下陷特征→炭疽病

10)黑褐色病斑,岭壳病害,不规则 ,无特征→炭疽病

2.1　数据项的建立

将病态的因素作为算法的数据项,由于每个病态因素又分为几个不同的情况,因此采用"病态因素.因素取值"的形式作为算法的最终数据项。结果如下:

1)病斑颜色:1.1,黑褐色病斑、1.2,粉红色病斑、1.3,褐色病斑

2)病斑位置:2.1,叶子病害、2.2,岭壳病害

3)病斑形状:3.1,圆形、3.2,半圆形、3.3,不规则

4)病斑特征:4.1,无特征、4.2,稍下陷特征、4.3,下陷特征

5)病斑名称:5.1,炭疽病、5.2,印度炭疽病、5.3,角斑病

根据经验数据得到初始数据库D(见表1)。

表1　初始数据库D

ID	数据项
001	1.1,2.1,3.1,4.1,5.1
002	1.1,2.1,3.1,4.1,5.1
003	1.2,2.2,3.2,4.2,5.1
004	1.3,2.1,3.1,4.1,5.2
005	1.3,2.1,3.1,4.1,5.2
006	1.3,2.2,3.1,4.3,5.3
007	1.1,2.2,3.3,4.1,5.3
008	1.1,2.1,3.1,4.1,5.1
009	1.2,2.2,3.2,4.2,5.1
010	1.1,2.2,3.3,4.1,5.1

2.2　连接规则的改动

经典的Apriori算法的连接规则是:对于两个频繁($k-1$)-项集,如果它们的前($k-2$)个项相同,就将它们进行连接。如对"1.1,2.1,3.1""1.1,2.1,3.2"连接得到"1.1,2.1,3.1,3.2"。而在建立专家系统的知识库时,由于"3.1"和"3.2"是同一种因素的两种不同的取值,它们两个是互斥的关系,不可能同时存在,因此对这种情况不进行连接。这样,连接之后,再经过最小支持度1的筛选,就会得到满足最小信任度与最小支持度阈值的频繁k-项集。对于上述农作物病害的例子,利用改动后的Apriori算法最终取得的为频繁5-项集,见表2。

表2　频繁5-项集

项集	1.1,2.1,3.1,4.1,5.1	1.1,2.2,3.3,4.1,5.1
支持度	3	1
项集	1.2,2.2,3.2,4.2,5.1	1.3,2.1,3.1,4.1,5.2
支持度	2	2
项集	1.1,2.2,3.3,4.1,5.3	1.3,2.2,3.1,4.3,5.3
支持度	1	1

2.3　规则生成方法的改动

在经典的关联规则算法中,是对频繁项集L产生它的所有非空子集,然后对L的每个非空子集S,如果满足S的信任度不小于最小信任度阈值,则产生一个关联规则"$s \rightarrow (l-s)$";

而在知识库的建立中,由于最后需要的是决策规则而不是关联规则,即所关注的重点不是几个属性之间的关联关系,而是由这几个属性组合后会得到什么样的结果。因此,只计算除结果之外的子集的信任度即可。如:对于频繁"1.2,2.2,3.2,4.2,5.1",只计算"1.2,2.2,3.2,4.2"的信任度,即S = 1.2,2.2,3.2,4.2,那么,对于满足信任度大于最小信任度阈值,产生$s \rightarrow (l-s)$的规则,即1.2,2.2,3.2,4.2→5.1。最后,再对照最初的属性值标号,转换成我们所理解的规则:粉红色病斑 ∧ 岭壳病害 ∧ 半圆形 ∧ 稍下陷特征 → 炭疽病。

对上述的农作物病虫害的例子生成的规则如下。

规则表:

1)黑褐色病斑∧岭壳病害∧不规则∧无特征→炭疽病　支持度:0.2,信任度:0.5

2)黑褐色病斑∧岭壳病害∧不规则∧无特征→角斑病　支持度:0.2,信任度:0.5

3)黑褐色病斑∧叶子病害∧圆形∧无特征→炭疽病　支持度:0.3,信任度:1

4)粉红色病斑∧岭壳病害∧半圆形∧稍下陷特征→炭疽病　支持度:0.2,信任度:1

5)褐色病斑∧叶子病害∧圆形∧无特征→印度炭疽病　支持度:0.2,信任度:1

6)褐色病斑∧岭壳病害∧圆形∧下陷特征→角斑病　支持度:0.1,信任度:1

(下转第80页)

1)建立一个连接和会话。

```
InitialContext ctx = newInitialContext();//获得 JNDI 上下文
QueueConnectionFactoryqcf = ctx. lookup ( connectionfactoryname);//获得连接工厂
Connection connection = qcf. createConnection();//获得一个连接
Session session = connection. createSession ( false, Session. AUTO_ACKNOWLEDGE);//从该连接获得一个会话
Destination dest1 = ( Queue ) jndiContext. lookup ("/jms/myQueue");//创建一个目的对象
```

2)创建 producer。

```
MessageProducer producer = session. createProducer (dest1);
```

3)发送消息。

需要传送的消息为征管数据库中的表的信息,例如税务、税种登记、消费税、企业所得税申报表等。发送前通过 Servlets 转换成序列化的对象。

```
Message m = session. createObjectMessage ();
Producer. send (m);
```

4)关闭 QueueConnection。

在程序块的最后一条语句是关闭连接。这一步很重要,忘记关闭 Connections 将导致服务器上的资源泄漏。

```
Connection. close ();
```

●JMS 接收程序(JMS Consumer)的实现(异步模式):即当消息发送者正在发送消息时,消息接受者无需处于运行状态。而是等接受者下次做好准备时,再将消息发送到接受者手上。

依据 JMS 对象模型,如图 2 所示,电子申报系统中的 JMS Consumer 客户端由下面的几个基本步骤来创建:

1)与 JMSProducer 相同:建立连接,创建会话。

2)创建 Consumer。

```
MessageConsumer consumer = session. createConsumer(dest1);
```

3)注册 listener。

```
MessageListener listener = newmyListener();
consumer. setMessageListener(listener);
```

4)调用 onMessage()方法。

```
public void onMessage ( Messagemsg ) {//readthemassageanddocomputation}
```

其中的 onMessage()方法需要由 listener 实现。

3 结束语

通过同步和异步通信的比较,根据网络电子申报的特点,选择了使用 JMS 技术。JMS 消息系统允许分开的未耦合的应用程序之间可靠地异步通信。对使用者,他不在乎谁产生消息,产生者是否仍在网络上以及消息是什么时候产生的。这就允许建立动态的、可靠的和灵活的系统。实践证明,JMS 技术在网络电子申报的应用是成功的。

参考文献:

[1] Gerton I, Almquist J, Cramer N. An Efficient, Scalable Content - Based Messaging System[A]. Seventh International Enterprise Distributed Object Computing Conference (EDOC'03)[C]. Crosspoint Blvd, Indianapolis: Wiley - IEEE Computer Society Press, 2003.

[2] Aleksy M, Schader M, Schnel A. Design and Implementation of a Bridge between CORBA Notification Service and the Java Message Service[A]. 36th Annual Hawaii International Conference on System Sciences (HICSS'03)[C]. Crosspoint Blvd, Indianapolis: Wiley - IEEE Computer Society Press, 2003.

[3] 罗晓斌,董守斌,徐 浩,等. 基于 JMS 的异步消息处理技术及应用[J]. 计算机工程,2002,28(12):121 - 122.

[4] Nei - Chiung P, Neung - Tsung T, Jen - Wei H. The Design and Implementation of A Real - Time Data Dispatching System[A]. Sixth IEEE International Symposium on Object - Oriented Real - Time Distributed Computing (ISORC'03)[C]. Crosspoint Blvd, Indianapolis: Wiley - IEEE Computer Society Press, 2003.

[5] 纪波林,王志坚. 基于 JMS 体系结构的消息服务技术的应用研究[J]. 计算机应用研究,2003(11):48 - 51.

(上接第 77 页)

这样,当农作物病害专家系统的用户将作物病害的几个特征输入后,系统就可以根据知识库中的规则表,将相应的规则输出。

3 结束语

专家系统着重解决的是推理决策问题,而数据挖掘中的关联规则则着重解决的是各个数据项之间有价值的联系。文中通过对经典关联规则的方法进行改动,找出了专家系统中各个因素的属性值组合与决策结果之间的关系,从而建立了专家系统的知识库。

参考文献:

[1] 黄梯云. 智能决策支持系统[M]. 北京:电子工业出版社, 2001. 103 - 128.

[2] 朱 明. 数据挖掘[M]. 合肥:中国科学技术大学出版社, 2002. 115 - 126.

[3] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[A]. Proceedings of the ACM SIGMOD Conference on Management of data[C]. Boston, MA, USA: [s. n.], 1993. 207 - 216.

[4] Agrawal R, Srikant R. Fast algorithms for mining association rules in large database[R]. Technical Report FJ9839. San Jose, CA: IBM Almaden Research Center, 1994.

[5] Aggarwal C, Agarawal R, Prasad V V V. Depth First Generation of Long Patterns[A]. In: The 6th ACM SIGKDD Intl Conf on Knowledge Discovery & Data Mining[C]. Boston, MA, USA: [s. n.], 2000.