关联规则挖掘算法在超市销售分析中的应用

唐 敏
(重庆工商大学管理学院 重庆 400067)

摘 要 销售数据分析是关联规则数据挖掘算法的主要应用领域之一,文章基于关联规则的算法理论,针对应用于超市销售关联规则的特点,提出了适用于超市销售相关性分析的模型。通过商业检验,该算法可以显著提高相关商品的销售额。

关键词 数据挖掘,关联规则,Apriori 算法,销售分析

Apply in the Sales Data Association Analysis with Association Analysis Data Mining Methodology

TANG Min
(Chongqing Technology and Business University, Chongqing 400067)

Abstract The sales data association analysis is one of the major application of the data mining methodology. Refering to the sales features of the retail points, the article sets up the retailing sales association analysis data model based on the association methodology. It proves that it can increase the sales dramatically by applying this methodology to the real business environment.

Keywords Data analysis, Association rule, Apriori methodology, Sales analysis

随着最终消费者消费水平的不断提高和生产市场由卖方市场向买方市场的转换,零售企业在物流供应链的重要意义不断凸显,已经成为优化供应链,提高整个社会生产效率的重要决定因素。如能挖掘出蕴含在海量销售数据中的经营技巧和市场规律,将优化零售企业产品品类配置,从而提升供应链上产业群的整体竞争力。关联规则作为数据挖掘的重要技术手段之一,在销售相关性分析中为提高零售企业产品配置合理性起到重要作用。

1 关联规则的基本概念

关联规则是从大量的数据中提取或"挖掘"出有用的知识,它能对过去的数据进行查询和遍历,找出过去数据之间的潜在联系,从而促进信息的显化。

1.1 关联规则的描述

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是项的集合,其中的元素称为项(item)。记 D 为交易 T 的集合,这里交易 T 是项的集合,并且 $T\subseteq I$。对应每一个交易有唯一的标识,如交易号(TID)。设 X 是一个 I 中项的一个集合,如果 $X\subseteq T$,那么称交易 T 包含 X。

一个关联规则是形如 $X\Rightarrow Y$ 的蕴涵式,这里 $X\subseteq I, Y\subset I$,并且 $X\cap Y=\Phi$。规则 $X\Rightarrow Y$ 在事物数据库 D 中的支持度(support)是事物集中包含 X 和 Y 的事物数与所有事物数之比,记为 support($X\Rightarrow Y$),即

$$\text{support}(X\Rightarrow Y)=|\{T;X\cup Y\subseteq T,T\in D\}|/|D|$$

规则 $X\Rightarrow Y$ 在事物集中的可信度(confidence)是指包含 X 和 Y 的事务数与包含 X 的事物数之比,记为 confidence($X\Rightarrow Y$),即:

$$\text{confidence}(X\Rightarrow Y)=|\{T;X\cup Y\subseteq T,T\in D\}|/|\{T:X\subseteq T,T\in D\}|。$$

给定一个交易集 D,挖掘关联规则问题就是产生支持度和可信度分别大于用户给定的最小支持度(minsupp)和最小可信度(minconf)的关联规则[1]。

1.2 Apriori 算法概述

Agrawal 等人于 1993 年首次提出布尔型关联规则问题,并给出了挖掘关联规则的算法 AIS,1994 年又提出了效率更高的 Apriori 算法[1~6]。该算法使用一种称作逐层搜索的迭代方法,需要多遍扫描事务数据库(D)。第一步扫描 D,对每个候选项计数 C_1,比较候选支持度计数与最小支持度计数,找出频繁 1 项集的集合 L_1。在接下的各步中,L_{k-1} 用于候选 C_k,如此下去,直到不能找到频繁 k-项集。Apriori 算法及以此为基础算法的缺点是要多遍扫描数据库并产生大量的候选项集。

针对 Apriori 算法框架的缺陷,Han 等人于 2000 年提出了 FP-Tree 结构和相应的算法,该算法无须生成候选 FP-Growth 项集,挖掘效率明显提高。但算法是通过逐步 FP-Growth 生成条件模式基和条件频繁模式树来挖掘频繁项集,因而影响了频繁项集的挖掘效率。于是 Wang 等人于 2002 年又提出了不需要生成条件模式基的自上而下挖掘的 FP-Tree 算法 TD-FP- Growth。

针对最小支持度难以恰当确定以及挖掘频繁模式和关联规则过多等问题,Han 等人又提出了 TFP 算法,该算法没有最小支持度的约束,运用了闭合模式挖掘长度大于 min_l、支持数最大的前 k 个频繁模式,得到了很好的结果。

2 超市销售关联规则应用

超市是零售企业最为普遍的业态之一,对于超市而言,不仅要关注销售产品品类间的联系,还要关注客户每次消费的销售总额及其分布特性,即需要在一个或多个属性上做聚合,

唐 敏 讲师,硕士。

只有当聚合的值高于指定的值时才做计算，该查询称为冰山查询[5]。Apriori 算法可以用来提高冰山查询的效率，其方法是先计算低维，只有当所有的低维都满足预制时才计算高维。通过 Apriori 算法对于销售数据进行关联挖掘需要进行数据准备、数据建模和模型评估几个过程。

2.1 数据准备

这个阶段的主要任务是收集数据(表、视图)，对可能无效的值(异常值、缺失值等)进行检查，并清理和格式化数据。在此还可以生成新的聚合属性，比如平均值和差。对于关联分析，该模型的输入表或者视图必须仅包含两个列，第一列是交易 id，而第二列是商品，它包含用来获得规则的元素。因此需要建立一个表或者视图，使得对于每笔交易，这个表中都有所购买的每件商品的交易 id 和分类 id，而不是该商品的商品 id。

本文采集分析数据来自于实际运作数据，共有 78,000 交易记录，为了防止错误数据产生，特意选定了一台 POS 机在一段时间稳定销售的销售数据，本文收集的数据中，包括的属性有：

“SALES_DATE”，“GOODS”，“SALES_TIME”，“PROD_ID”，“BAR”，“UNIT”，“QTY”，“PRICE”，“NET_AMOUNT”，“DISCOUNT_AMOUNT”，“PRICOST”，“SPLIT”，“VIP_NO”，“SERIALNO”，“MACHINE”，“SALES_GROUP”，“JOB”，“SALER_NAME”，“STORE_CODE”，“EDI_SENT_DATE”，“CREATED_BY”，“CREATED_DATE”。

对于实际的数据挖掘而言，只有 Sales_Item 和 Goods 这两个域有用，为了表示出在一次交易中的所有销售商品，特规定在同一时间的商品算是同一次交易的商品。因此，初步清洗之后的信息就只包括“SALES_TIME”和“GOODS”两项。

2.2 数据建模

本文使用 IBM Visualization 作为挖掘工具，作为初始步骤，将可信度 Confidence 设置为 25%，将支持度 Support 设置为 0.5%。

```
[tutorial_path]:\>db2 -tvf retail_assoc_view.db2
connect to RETAIL;
call IDMMX.BuildRuleModel('RETAIL.ASSOC_MODEL',
'RETAIL.ASSOC_VIEW',
'TRANSID',0.5,25,3,
'DM_addNmp("NewMap","RETAIL.ARTICLE_CATEGORIES","ID","DESC"),
DM_setFldNmp("CATEGORYID","NewMap")');
connect reset;
```

2.3 模型评估

模型的评估就是凭评估该模型作为结果来生成的规则，该任务是选取哪些规则提供了对解决业务问题有用的知识，在这一步中必须与业务专家合作分析，根据和超市销售主管共同分析，得出了相关的规则，举例如下：

例 1

[徐福记 DODO 糖 38g 水蜜桃]⇒[徐福记 DODO 糖 38g 乳酸]
Support=0.1011% Confidence=22.1700% Lift=46.1095

这个规则解释如下，购买了徐福记 DODO 糖 38g 水蜜桃商品的顾客在 22.1700%的情况下也会购买徐福记 DODO 糖 38g 乳酸。这条规则在商业运作是合理的，因为这两种商品是散装商品并且通常情况下都会被摆放在同一个区域，客户很容易在购买水蜜桃 DODO 糖时也会顺手拿一些乳酸 DODO 糖。

例 2

[散装小食]+[洽洽香瓜子 380g]⇒[徐福记散装布丁、果冻]
Support=0.1078% Confidence=51.6100% Lift=8.5839

购买散装小食和洽洽香瓜子 380g 的顾客，在 51.6100%的情况下也会购买徐福记散装布丁、果冻。这条规则影响总交易量的 0.1078%。此外，在具有散装小食和洽洽香瓜子 380g 的那些交易中找到徐福记散装布丁、果冻的概率，是在其他所有交易中找到它们(徐福记散装布丁、果冻)的概率的 8.5839 倍。

除了在图表中已经生成的关联产品，还可以生成向顾客推荐的其他产品，为避免给新顾客增添烦恼和压力，可以仅推荐具有较高 lift 或 support 值的前三种产品。

3 商业检验

根据以上分析数据，与商业伙伴共同制定了如下的计划去实施关联规则的实际应用：

a. 针对分析中显示关联性很高的商品，调整其商品的陈列，把一些销售状况不好的商品剔出，将相关性高的商品陈列在一块进行捆绑销售；

b. 针对其他关联性商品，与市场部制定了市场促销计划。在销售促销单明确而且显眼的地方，把关联性商品对消费者进行推荐。

整个计划实施了一个月，促销是按照每两周一次的方式进行。整个计划得到了很大的成功，进行实施的单店的整体销售额同比去年相同时期增长了 20%，相关商品的销售额不同程度地增长了 40%～60%。

结论 本文只是对零售业中普通的超市进行数据挖掘进行研究，就零售行业而言，还有便利店、百货店以及购物中心等其他重要的业态没有进行研究。对整个零售行业而言，超市所占的销售份额是较小的，如果能够把这种方法和理念有效地推广到其他业态，将对零售业的销售业绩提升产生积极的影响。

参 考 文 献

1 Agrawal R, Imielinski T, Swami A. Mining Association Rules-Between Sets of Items in Large Databases [J]. In: Proc. of ACM SIGMOD Conference on Management of Data, 1993. 207～216

2 Agrawal R, Srikant R. Fast Algorithms for Mining Association-Rules [J]. In: Proc. of International Conference on VeryLarge Databases, 1994. 487～499

3 Han J, Pei J, Yin Y. Mining Frequent Patterns Without Candidate Generation. In: Proc. of ACM-SIGMOD Conference, 2000. 1～12

4 Wang K, Tang L, Han J, et al. Top Down FP-growth for Association Rule Mining. In: Proc. of PAKDD2002, 2002. 334～340

5 Han J, JianYong, Lu Y P, et al. Mining Top-K Frequent Closed Patterns Without Minimum Support. In: ICDM, 2002. 211～218

6 闫莺，王大玲，于戈. 支持个性化推荐的 Web 页面关联规则挖掘算法. 计算机工程，2005，31(1)：79～81