基于外部环境的关联规则挖掘

万福才[1],唐明慧[2],陈 晓[3]
(1. 沈阳大学 信息工程学院,辽宁 沈阳 110044;
2. 沈阳大学 科技中心,辽宁 沈阳 110044;
3. 华东交通大学 信息工程学院,江西 南昌 330013)

摘 要:传统的关联规则挖掘不能发现具有潜在价值的关联规则,如在挖掘交易数据库时,一些包含新商品的关联规则往往由于其信任度低而被删除,但是外部环境的动态性使得这些规则在某些特定时期对用户有很大的价值性。为了解决这个问题,保留具有潜在价值的关联规则,文中提出了基于外部环境的关联规则数据挖掘方法。在挖掘过程中,重新定义了信任度,并提出经济价值度的概念,根据信任度和经济价值度,可以有效地实现关联规则冗余性大小的排序,保留具有潜在价值的关联规则,适应用户的需求。实验表明,该方法可以有效地保留具有潜在价值的关联规则。

关键词:关联规则;信任度;经济价值度;潜在价值
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2013)01-0115-04
doi:10.3969/j.jssn.1673-629X.2003.01.029

Association Rule Mining Based on External Environment

WAN Fu-cai[1], TANG Ming-hui[2], CHEN Xiao[3]
(1. College of Information Engineering, Shenyang University, Shenyang 110044, China;
2. Scientific Technology Research Center, Shenyang University, Shenyang 110044, China;
3. School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: The association rules that have potential value can't be found by traditional association rules mining. For example, when in the transaction database mining, some association rules that contain new goods information have always been removed, because of its low confidence. But these rules may have potential value. Users can't get these deleted rules. But if the external environment changed, these rules may be useful to users. In order to solve this problem and reserve the rules contain potential value, put forward a new method named association rules mining based on the external environment. Redefined the confidence in the proposed method, and put forward the degree of economic value. According to this method can reserve the rules contain potential value and fulfill the ranking of association rules, then meet the demand of user. The experimental results show that the method reserves the rules contain potential value more efficiently.

Key words: association rule; confidence; the degree of economic value; potential value

0 引 言

随着数据挖掘技术的发展,关联规则的应用也越来越广泛,典型的关联规则发现问题是对超市中的货篮数据进行分析[1~3],通过挖掘海量商业记录发现的关联规则反映了不同商品之间关系,通过分析顾客的购买行为,便于商家及时调整商品分类、降价经销、货架摆放、绑定促销等策略,提高营业额,为用户创造经济价值。传统的关联规则挖掘预先设定最小支持度阈值和最小信任度阈值,只有满足这两个阈值的关联规则才会呈现给用户。但是针对商业应用这一特定领域,这种传统方式有很大的缺陷。首先,挖掘出的关联规则中存在冗余规则。许多学者都针对这个问题进行了研究,但是大都采用基于一定的约束采用某种技术或算法删除冗余规则的方法[4~11],但是都无法确定删除的规则对用户永远都没有价值。其次,传统的关联规则挖掘根据规则信任度来判断是否有价值,如果信任度不小于最小信任度阈值则认为值得保留否则就舍弃,这种方式很容易丢掉具有潜在价值的规则。例如用传统的方式挖掘交易数据库时,关联规则 $A \Rightarrow B$,由于其信任度很低认为没有价值而被删除了。但是由于没有考虑到 B 是一种很有竞争力的新商品,因而失去了一次商机。

收稿日期:2012-05-16;**修回日期**:2012-08-22
基金项目:辽宁省自然科学基金项目(20102153)
作者简介:万福才(1967-),男,辽宁沈阳人,博士,教授,研究方向为数据挖掘、智能优化算法、新产品投入、电子商务建模与优化;唐明慧(1988-),女,山东临沂人,硕士研究生,研究方向为电子商务与商务智能。

针对这两个问题,提出了一种解决的方法,基于外部环境的关联规则挖掘,重新定义了信任度和引入经济价值度。信任度在传统信任度定义的基础上加入了保留权重和新商品的频度,这个定义是针对包含新商品这种特殊关联规则的,保留权重和新商品的频度越大,规则的信任度就越大,保留的可能也就越大。这样一来就可以避免丢失具有潜在价值的规则。构造规则的经济价值度时主要考虑信任度和规则前后件的重要程度之比,重要程度用当前市场价格来衡量。经济价值度越大表示预计规则产生的经济效益越大,成为冗余规则的可能性越小。根据经济价值度对关联规则的进行排序,用户根据排序选择关联规则。

1 相关定义

定义 1:潜在价值的信息是指当前没有体现出很大价值但是在未来一段很短的时间内一定会体现出很大价值的信息。

潜在价值的关联规则是针对那些规则前件或后件中包含新商品的特殊规则而言的,如果特殊规则满足当前对用户没有很大价值,但是在很短时间后就可能会给用户带来很大价值这一条件,即为潜在价值的关联规则。

定义 2:设 $N=\{n_1,n_2,\cdots,n_k,\cdots,n_m\}$ 为事物数据库 D 中所有新商品的集合,称为新项集。设用户给定的最小支持度和最小信任度分别为 c 和 s,则对于关联规则 $A\Rightarrow B$,其信任度记为 confidence $(A\Rightarrow B)$:

(1)当项集 A 与 B 满足:$A\cap N=\varnothing$ 且 $B\cap N=\varnothing$

$$\text{confidence}(A\Rightarrow B)=\frac{\text{support}(A\cup B)}{\text{support}(A)}$$

当 confidence $(A\Rightarrow B)\geqslant s$ 时,关联规则 $A\Rightarrow B$ 会输出,呈现给用户。

(2)当项集 A 与 B 满足:$A\cap N\neq\varnothing$,$B\cap N\neq\varnothing$ 至少有一个成立时

$$\text{confidence}(A\Rightarrow B)=\frac{\text{support}(A\cup B)}{\text{support}(A)}+\overline{\omega}s+M$$

其中 s 为用户给定的最小信任度,$\omega=\{\omega_1,\omega_2,\cdots,\omega_k,\cdots,\omega_m\}$ 为 $N=\{n_1,n_2,\cdots,n_k,\cdots,n_m\}$ 中对应商品的权重,权重越大表示这种新商品的发展潜力越大,权重可以随着外部环境的变化而进行相应的调整。$\overline{\omega}$ 为保留权重,即 A 与 B 中包含的所有新商品的权重的平均值。M 为关联规则 $A\Rightarrow B$ 中新商品的频度:

$$M=\frac{\text{support}(X)}{\max\ \sup}s$$

其中 $X=A\cap N\cup B\cap N$,max sup 为频繁项集支持度的最大值。

当 confidence $(A\Rightarrow B)\geqslant s$ 时,关联规则 $A\Rightarrow B$ 即为潜在价值的关联规则,对于潜在价值的关联规则也会输出,呈现给用户。

定义 3:一个项集的价值为所有组成元素的价值之和,记为 F,元素的价值用商品的市场价格来衡量,记为 f,两个项集的价值之比称为价值区分度,记为 λ,设项集 $A=\{A_1,A_2,\cdots,A_m\}$,$B=\{B_1,B_2,\cdots,B_n\}$,则项集 A 与 B 的价值区分度为:

$$\lambda=\frac{F_A}{F_B}=\frac{\sum_{i=1}^{m}f_i}{\sum_{j=1}^{n}f_j}$$

经济价值度表示关联规则预计给用户带来经济收益的大小程度,关联规则 $A\Rightarrow B$ 的经济价值度,记为 value $(A\Rightarrow B)$,则:

$$\text{value}(A\Rightarrow B)=\text{confidence}(A\Rightarrow B)+\text{sgn}(x)\,\mu\,e^{\lambda}$$

其中 μ 为平衡因子,$\mu\in[0,1]$,λ 为项集 A 与 B 的价值区分度,$x=F_A-F_B$

$$\text{sgn}(x)=\begin{cases}1 & x\geqslant 0\\ -1 & x<0\end{cases}$$

2 基于外部环境的关联规则数据挖掘过程描述

基于外部环境的关联规则数据挖掘与现今存在的关联规则数据挖掘不同之处在于,在挖掘的过程中考虑了外部环境的影响,具有潜在价值的关联规则也会呈现给用户,降低因环境剧烈变动导致损失的可能性,同时用经济价值度为衡量尺度,将挖掘出的关联规则排序。经济价值度越大,排列顺序越靠前。这种方式方便用户根据价值大小取舍关联规则,降低因环境剧烈变动导致损失的可能性。

目前有多种针对关联规则的挖掘算法,较典型的有 Apriori,FP 树以及基于 Apriori 的改进算法[12,13]。

基于外部环境的关联规则数据挖掘需要经历如下 3 个步骤:

(1) 找出所有的频繁项集。

这一过程可以通过 Apriori 算法实现,通过构造潜在频繁项集,检索出事务数据库中所有的频繁项集,当然也可以根据事务数据库的特点采用其他算法。

(2) 找出所有满足最小信任度阈值的关联规则。

这一过程首先要扫描事务数据库,检索出所有的新商品,构造出新项集 $N=\{n_1,n_2,\cdots,n_k,\cdots,n_m\}$,然后根据条件计算出关联规则的信任度。

当关联规则 $A\Rightarrow B$ 满足:$A\cap N=\varnothing$ 且 $B\cap N=\varnothing$

$$\text{confidence}(A\Rightarrow B)=\frac{\text{support}(A\cup B)}{\text{support}(A)}$$

当关联规则 $A\Rightarrow B$ 满足:$A\cap N\neq\varnothing$,$B\cap N\neq\varnothing$ 至

少有一个成立时

$$\text{confidence}(A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)} + \bar{\omega}s + M$$

权重随外部环境而调整,信任度也就随之变化。计算出关联规则的信任度之后,与用户给定的最小信任度阈值做比较。

(3)计算满足最小信任度阈值的关联规则的经济价值度,并根据经济价值度大小对关联规则进行排序。

如果关联规则 $A \Rightarrow B$ 满足:$\text{confidence}(A \Rightarrow B) \geqslant s$,则根据经济价值度定义计算出 $\text{value}(A \Rightarrow B)$。计算出所有满足条件的关联规则的经济价值度,然后按照经济价值度大小对关联规则进行排序,排序后的关联规则呈现给用户。

3　实验结果

实验编程环境选用 Visual Studio 2005 ,C#语言。实验数据集随机生成,最小支持度设为 2% ,最小信任度设为 10% ,实验用到的权重在 0 ~ 1 之间取值。

采用传统的关联规则挖掘和基于外部环境的关联规则挖掘两种方法,对同一事物数据库进行关联规则的挖掘,对比同一规则在使用两种不同方法时的信任度。文中选取 103 条关联规则(包含新商品的关联规则全部选取)进行结果分析。结果如图 1 所示:

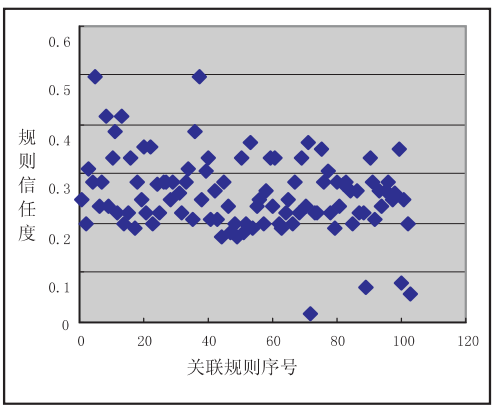

图 1　传统关联规则挖掘

图 1 是采用传统关联规则挖掘方法,图 2 采用基于外部环境的关联规则挖掘方法,由图 1 可以看出在选取的 103 条关联规则中有 4 条包含新商品的关联规则信任度小于 10% ,采用基于外部环境的关联规则挖掘方法时,可有效提高具有潜在价值的关联规则的信任度,由图 2 可看出所有规则的信任度均不小于 10% 。

图 3 传统挖掘与基于外部环境挖掘的结果曲线对比分析图,从图中可以看出具有潜在价值的关联规则的信任度均有不同程度的提高,降低了具有潜在价值的关联规则因为信任度低而被舍弃的概率,实现了保留具有潜在价值的关联规则的目的。

从实验结果可以看出,采用基于外部环境的关联规则挖掘方法是可行的,可以明显提高具有潜在价值的关联规则的置信度,达到预期的目的。

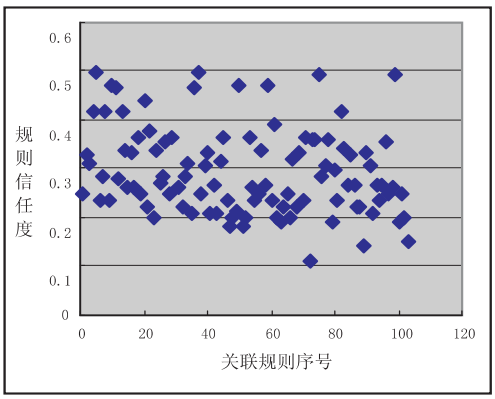

图 2　基于外部环境的关联规则挖掘

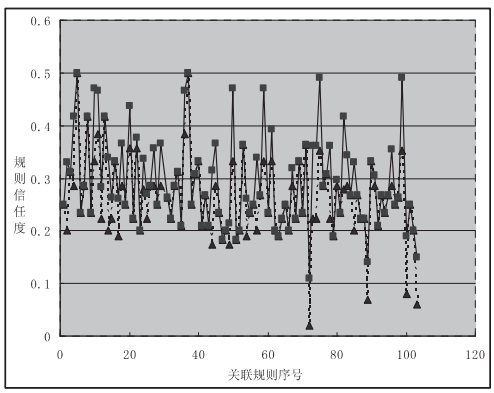

图 3　传统挖掘与基于外部环境挖掘的结果对比

(三角形/正方形)

采用基于外部环境的关联规则挖掘将满足最小信任度的关联规则呈现给用户,如果规则的信任度差异很小的话,对用户的选择会造成困扰,如图 2 所示的 103 条关联规则信任度比较集中,不利于用户区分选择。所以文中进一步计算了这些关联规则的经济价值度($\mu = 0.6$),结果如图 4 所示:

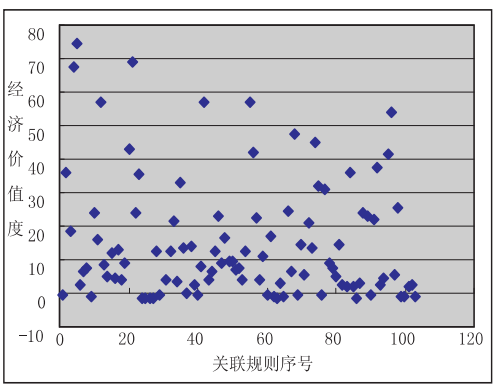

图 4　经济价值度

经济价值度是从实际经济价值的角度,对关联规则进行区分,更贴近用户使用关联规则创造经济价值的目的。与图 2 相比,图 4 中规则的分散程度变大了,区分度也提高了,用户可以按照经济价值的大小来选择规则。从实验结果可以看出,使用经济价值度来选

择关联规则是一种行之有效的方法,这种选择方式更合理、更容易。

外部环境的变化会导致保留权重和价值区分度发生相应的变化,用户可以及时提供给数据挖掘相关人员,及时的调整参数,重新生成与外部环境相适应的关联规则,用经济价值度进行排序,方便用户从经济价值的角度进行取舍。

4 结束语

基于外部环境的关联规则挖掘重新定义信任度,实现保留具有潜在价值的关联规则的目的,并且根据经济价值度对关联规则进行排序,为用户取舍关联规则提供依据,降低因环境剧烈变动导致损失的可能性。

这种挖掘方式考虑了外部环境的变化和用户的经济效益,有很大的现实意义,但也有不足之处,例如保留权重是人为设定的,容易产生误差,进而影响结果,所以对于如何更客观的设定权重,得到有价值的关联规则,还需要进一步的深入研究。

参考文献:

[1] 谭 华,谢 赤,储慧斌. 基于模糊关联规则的股票市场交易规则抽取[J]. 系统工程,2007,25(4):92-97.

[2] 井福荣,谢辅雯. 关联规则在网站结构优化中的改进算法[J]. 计算机系统应用,2007(1):43-46.

[3] 李雄飞,李 军. 数据挖掘与知识发现[M]. 北京:高等教育出版社,2003.

[4] Li G,Hamilton H J. Basic association rules[C]//Proceedings of the 4th SIAM International Conference on Data Mining. Orlando,USA:[s. n.],2004:166-177.

[5] Bastide Y,Pasquier N,Taouil R,et al. Mining minimal non-redundant association rules using frequent closed itemsets[C]//Proceedings of the 1st International Conference on Computational Logic. Berlin,German:Springer,2000:972-986.

[6] Xu Yue,Li Yuefeng. Mining non-redundant association rules based on concise bases[J]. International Journal of Pattern Recognition and Artificial Intelligence,2007,21(4):659-675.

[7] Loglisci C,Malerba D. Mining multiple level non-redundant association rules through two-fold pruning of redundancies[C]//Proceedings of MLDM. [s. l.]:[s. n.],2009:251-265.

[8] Cheng J,Ke Y P,Ng W. Effective elimination of redundant association rules[J]. Data mining and knowledge discovery,2008,16(2):221-249.

[9] 陈 茵,闪四清,刘 鲁,等. 最小冗余的无损关联规则集表述[J]. 自动化学报,2008,34(12):1-7.

[10] 李 帆,夏士雄,张 磊,等. 基于语义划分的多层次关联规则冗余处理方法[J]. 微电子学与计算机,2010,27(8):252-260.

[11] 韦素云,吉根林,曲维光. 关联规则的冗余删除与聚类[J]. 小型微电子计算机系统,2006,27(1):110-113.

[12] 文 拯,梁建武,陈 英. 关联规则算法研究[J]. 计算机技术与发展,2009,19(5):56-58.

[13] 张广路,雷景生,吴兴惠. 一种改进的 Apriori 关联规则挖掘算法[J]. 计算机技术与发展,2010,20(6):84-88.

(上接第 114 页)

向量空间模型中包含非常多的 0 值,即该向量空间模型是一种稀疏矩阵,因此在后续的研究工作中会重点关注特征的类别区分能力度量以及特征数据稀疏处理的问题。

参考文献:

[1] 赵妍妍,秦 兵,刘 挺. 文本情感分析[J]. 软件学报,2010,21(8):1834-1848.

[2] Salton G,Wang A,Yang C S. A vector space model for automatic indexing[J]. Communication of the ACM,1975,18(11):613-620.

[3] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques[C]//Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing. [s. l.]:[s. n.],2002:79-86.

[4] 徐 军,丁宇新,王晓龙. 使用机器学习方法进行新闻的情感自动分类[J]. 中文信息学报,2007,21(6):95-100.

[5] 周 杰. 基于机器学习的网络新闻评论情感分类研究[J]. 计算机应用,2010,30(4):1011-1014.

[6] Cevikalp H,Neamtu M,Wilkes M. Discriminative Common Vector Method with Kernels[J]. IEEE Transactions on Neural Networks,2006,17(6):1550-1565.

[7] Mikat S,Fitscht G,Weston J. Fisher discriminant analysis with kernels[C]//Neural Networks for Signal Processing IX. New York:IEEE,1999:41-48.

[8] 眷红英,郭 明,柴玉梅,等. 新闻报道文本的情感倾向性研究[J]. 计算机工程,2010,36(15):20-22.

[9] Deng Weihong,Hu Jiani,Guo Jun. Robust Fisher Linear Discriminant Model for Dimensionality Reduction[C]//International Conference on Pattern Recognition (ICPR 2006). [s. l.]:[s. n.],2006:699-702.

[10] 徐庆伶,汪西莉. 一种基于支持向量机的半监督分类方法[J]. 计算机技术与发展,2011,21(10):115-117.

[11] Yang Shu,Yan Shuicheng,Zhang Chao. Bilinear Analysis for Kernel Selection and Nonlinear Feature Extraction[J]. IEEE Transactions on Neural Networks,2007,18(5):1442-1452.

[12] 徐淑坦. 基于改进 RBF 神经网络的文本情感分类研究[D]. 长春:吉林大学,2011.

基于外部环境的关联规则挖掘

作者： 万福才， 唐明慧， 陈晓
作者单位： 万福才(沈阳大学 信息工程学院,辽宁 沈阳 110044)， 唐明慧(沈阳大学 科技中心,辽宁 沈阳 110044)， 陈晓(华东交通大学 信息工程学院,江西 南昌 330013)

刊名： 计算机技术与发展
英文刊名： Computer Technology and Development
年，卷(期)： 2013(1)