超越支持度-置信度框架的负相关对规则挖掘*)

钱铁云[1] 冯小年[2] 王元珍[1]
(华中科技大学计算机学院数据库与多媒体技术研究所 武汉 430074)[1]
(中国电力财务有限公司华中分公司 武汉 430062)[2]

摘 要 相关规则比传统的关联规则更具有实际意义。但现存的相关规则挖掘算法均需利用 apriori 类似算法挖掘具有高支持度的项集,再对获得的项集进行相关性测试而获取相关规则,这导致低支持度-高相关度的规则不易被发现。直接挖掘相关规则的困难在于候选相关项不能利用 apriori 类似性质进行剪枝,导致搜索空间爆炸性增长。本文提出的算法 MNI 利用 Phi 相关系数的下界来产生候选负相关项,从而缩小负相关项搜索空间,并证明了该算法的完全性和正确性。在负相关项对基础上利用规则可靠度产生负相关规则时,提出将负相关对计数统一转化为正相关对计数的方法。在真实数据集上的实验结果表明,该算法 MNI 能有效提高负相关项对的挖掘速度。

关键词 关联规则,相关规则,Phi 相关系数,规则可靠度

Mining Negative Correlation Rules Beyond Support-Confidence Framework

QIAN Tie-Yun[1] FENG Xiao-Nian[2] WANG Yuan-Zhen[1]
(Computer Science Department, Huazhong University of Science and Technology, Wuhan 430074)[1]
(China Power Finance Company, Huazhong Branch, Wuhan 430062)[2]

Abstract High correlation rules are more practical than traditional association rules, but existed correlation rule mining algorithms are almost apriori-based. This results in the difficulty of finding correlation rules with low support but high correlation. In this paper a new algorithm called MNI is introduced to use the lower bound of Phi correlation coefficient to generate all candidate negative correlation items and reduce explosive search space. Both the completeness and correctness of MNI are proved. Negative correlation rules are mined using reliability measure without directly counting the number of negative correlation pairs. Experiments on real datasets show that the algorithm is quite efficient in negative correlation items mining.

Keywords Association rules, Correlation rules, Phi correlation coefficient, Rule reliability measure

1 引言

关联规则挖掘用于寻找大型事务数据库中项之间的有趣关系。针对该问题已经进行了广泛的研究,其中最为著名的算法是 apriori 算法[1]和 fp-tree 方法[2]。但是,使用支持度-置信度框架的关联规则挖掘均存在以下问题:尽管 $A=>B$ 是符合最小支持度、置信度要求的强规则,但是 A 和 B 的出现是独立的(或基本独立),即它们之间实际上并没有(或很少有)相关性。相关性问题首先在文[3]中被研究,该文提出利用 χ^2 进行显著性测试,再利用 $P(A \wedge B)/(P(A) * P(B))$ 来判断 A 与 B 之间存在的正、负相关性。A、B 间支持度和相关性高低程度可能存在如下组合(按支持度-相关性):Low-Low、High-Low、High-High、Low-High。L-L 型显然不属于问题考虑范围,H-L 型则存在如前所述的缺点,H-H 型是必须考虑的,L-H 型在事务数据库中可用于挖掘较少购买但是非常昂贵物品如项链和耳环之间的关联关系,而在关联文本分类中则可以用来挖掘稀有的但是非常专用的词和类别如利福平和医药类之间的关系。因此挖掘高相关性规则(H-H 和 L-H 型)比挖掘高支持度规则更具有实际意义。

支持度-置信度框架算法首先在 k-1 频繁项的基础上产生 k-候选频繁项,然后利用 apriori 性质对候选项进行剪枝最后统计计数以产生 k-频繁项。该过程迭代直到候选项为空。在频繁项集基础上再利用置信度获取规则。具有该结构的算法被称为是 Apriori-Like 算法,但是该类算法只能用于挖掘 H-H 或 H-L 型的规则,而不能用于挖掘 L-H 型的规则,因为一旦把最小支持度阈值设得很低,则剪枝的有效性就无法体现,频繁项的爆炸增长使得挖掘算法效率极为低下乃至根本不可行。

现存考虑规则相关性的挖掘算法基本上属于 Apriori-Like 算法的改进,一般是在挖掘出的 k-频繁项基础上加上某种相关性测度。如文[3]寻求 p-支持度定义下的频繁项,文[4]在 k-频繁项上结合领域知识利用规则兴趣度挖掘负关联规则。文[5]讨论了当相依表大于 2×2 时 χ^2 统计存在的问题并提出一种新的关联规则可靠性度量,但是并没有提出实际的算法。文[6]方法稍有不同,其正规则的求解利用传统的 k-频繁项的基础上加入最小兴趣度度量的测试获得,负规则求解方法是通过{k-候选频繁项}-{k-频繁项}获取 k-非频繁项,但是 k-非频繁项的任意 k-1 子项仍然是频繁的,因此算

*)本文研究获高等学校博士学科点专项科研基金:基于浓缩数据立方的联机分析处理与梯度挖掘(项目编号 20030487032)资助。钱铁云 博士研究生,研究方向为数据挖掘和信息检索。冯小年 硕士,主要从事现代数据库研究。王元珍 教授,博士生导师,主要从事现代数据库研究。

法仍然基于 Support-Confidence 框架结构。文[7]求的是后件确定的相关规则，方法是在 k-候选频繁项基础上通过相关系数验证来求相关项。

相关项并不呈现 apriori 类似性质，即任何相关项的非空子集并不一定相关。但若不能利用 apriori 性质剪枝，那么求相关规则的困难在于搜索空间的爆炸增长使得 Naïve 算法根本不可行。文[8]中的算法是唯一非 Apriori-Like 的，它讨论了针对全部相关项对的查询请求并提出利用 Phi 相关系数的上界来缩小搜索空间，但文章只设计了求正相关项对的算法。在事务数据库中，负相关规则表示顾客购买了某些商品就会导致不购买另外一些商品，例如 68%买了咖啡的顾客就不再购买茶叶。负相关规则描述了数据库中存在的反常事件，与正相关规则具有同样重要的意义。相对于正相关规则，负相关规则挖掘更为困难，其原因在于事务数据库中只保存顾客购买了什么商品，而不会保存顾客没有购买什么商品，这使得负项集的计数非常困难。

本文旨在利用 Phi 相关系数的下界缩小负相关项对的搜索空间，从而只需在内存中保存候选的负相关项对，再通过求实际的相关系数确定真正的负相关项对，在利用可靠度度量从相关项求负相关规则时，针对直接进行负相关项对计数困难的问题，提出一种将负相关项对计数统一转化为求正相关项对计数的方法。

本文组织如下：第 2 部分给出相关概念，第 3 部分介绍负相关规则的挖掘，第 4 部分对算法的完全性和正确性进行证明，第 5 部分给出实验结果，最后为结论和将来工作提示。

2 相关概念

假设 $I=\{I_1, I_2, \cdots, I_m\}$ 是 m 个不同项 (items)的集合。给定一个事务数据库 D，其中每个事务 T(transaction)是 I 中一组项目的集合，即 $T\subseteq I$。在本文中，仅考虑项对 A、B 之间的关系，$A\in I$，$B\in I$。

表 1 Table1 二维变量相依表

	B	¬B	Row Total
A	f11	f10	f1+
¬A	f01	f00	f0+
Column Total	f+1	f+0	N

项相关度(Item Correlation) 对项对(A,B)，其相关度定义为：

$$\phi_{A,B}=\frac{f_{00}f_{11}-f_{01}f_{10}}{\sqrt{f_{+0}f_{+1}f_{0+}f_{1+}}}$$

项支持度(Item Support) 项 A 的支持度指 D 中事务包含项 A 百分比，$s(A)=P(A)$；项对(A,B)的支持度指 D 中事务包含项 A 和 B 两者的百分比，$s(A,B)=P(A,B)$。

规则可靠度(Rule Reliability) 对规则 $A=>B$，可靠度反映了 A 出现情况下 B 出现的条件概率和 B 单独出现的概率之间的差距，$r(A=>B)=P(B|A)-P(B)$。

正相关规则(Positive Correlation Rule) 定义 $A=>B$ 为正相关规则，若①项对(A,B)是正相关的，即 $\phi_{A,B}\geqslant$ min-corr；②规则 $A=>B$ 是可靠的，即 $r(A=>B)\geqslant$min-reli。

负相关规则(Negative Correlation Rule) 定义 $A=>\neg B$ 为负相关规则，若①项对(A,B)是负相关的，即 $\phi_{A,B}\leqslant-$ min-corr；②规则 $A=>\neg B$ 是可靠的，即 $r(A=>\neg B)\geqslant$ min-reli。

上述定义中的 min-corr 和 min-reli 分别为用户指定的最小相关度、最小可靠度，且 min-corr>0，min-reli >0。

3 负相关规则对挖掘

3.1 ϕ相关系数的下界

考察相关系数 $\phi_{A,B}$，并把它稍作变形，可得：

$$\phi_{A,B}=\frac{f_{00}f_{11}-f_{01}f_{10}}{\sqrt{f_{+0}f_{+1}f_{0+}f_{1+}}}=\frac{(N-f_{01}-f_{10}-f_{11})f_{11}-f_{01}f_{10}}{\sqrt{f_{+0}f_{+1}f_{0+}f_{1+}}}$$

$$=\frac{\frac{f_{11}}{N}-\frac{f_{+1}f_{1+}}{N\ N}}{\sqrt{\frac{f_{+0}}{N}\frac{f_{+1}}{N}\frac{f_{0+}}{N}\frac{f_{1+}}{N}}}=\frac{s(A,B)-s(A)s(B)}{\sqrt{s(A)s(B)(1-s(A))(1-s(B))}} \quad (1)$$

文[8]中已经证明了该相关系数存在上界，并可以用来裁剪正相关项对的搜索空间。本文将证明该相关系数存在下界，并可以用来裁剪负相关项对的搜索空间。

首先考虑边界点：$s(A)$、$s(B)$取{0,1}的时候，式(1)的分母为 0，但分子也为 0(见式(2))。考虑到分子 $s(A,B)-s(A)*s(B)$反映了 A,B 一起出现的概率和 A 和 B 单独出现的概率之差异，而公式的分母其实是起到一个规范化的作用，那么在边界点处的两个项是互相独立的，可以不必考虑。

$$s(A,B)-s(A)*s(B)=$$
$$\begin{cases} s(A,B)-0=0-0=0, s(A)ors(B)=0 \\ s(A,B)-s(B)=s(B)-s(B)=0, s(A)=1 \\ s(A,B)-s(A)=s(A)-s(A)=0, s(B)=1 \end{cases} \quad (2)$$

定理 1 ϕ相关系数的下界在 $s(A,B)=0$ 时取得，且

$$\text{lower}(\phi_{A,B})=-\frac{\sqrt{s(A)s(B)}}{\sqrt{(1-s(A))(1-s(B))}} \quad (3)$$

证明：根据概率知识，我们知道若 A、B 存在正相关，则 $s(A,B)>s(A)s(B)$，$\phi_{A,B}>0$；反之若 A、B 存在负相关，则 $s(A,B)<s(A)s(B)$，$\phi_{A,B}<0$；若 A、B 不相关，则 $s(A,B)=s(A)s(B)$，$\phi_{A,B}=0$。

由于 $0\leqslant s(A,B)\leqslant\min(s(A),s(B))$

$$\phi_{A,B}\geqslant\frac{0-s(A)s(B)}{\sqrt{s(A)s(B)(1-s(A))(1-s(B))}}$$

$$=-\frac{\sqrt{s(A)s(B)}}{\sqrt{(1-s(A))(1-s(B))}}$$

3.2 ϕ相关系数下界的条件单调性及应用

定理 2 在式(3)中，如果 $s(A)$固定，则 $\text{lower}(\phi_{A,B})$随 $s(B)$的增大而减小。

证明：$\text{lower}(\phi_{A,B})=-\frac{\sqrt{s(A)s(B)}}{\sqrt{(1-s(A))(1-s(B))}}=-\frac{\sqrt{s(A)}}{\sqrt{1-s(A)}}*\frac{1}{\sqrt{\frac{1}{s(B)}-1}}$

任意项 B1、B2 的支持度满足以下不等式：$0\leqslant s(B1), s(B2)\leqslant 1$，

若 $s(B1)>s(B2)$，有

$$\frac{1}{\sqrt{\frac{1}{s(B1)}-1}}>\frac{1}{\sqrt{\frac{1}{s(B2)}-1}}$$

则对相同的 $s(A)$，$\text{lower}(\phi_{A,B1})<\text{lower}(\phi_{A,B2})$

定理 3 给定 $s(B1)>s(B2)$，则如果 $\text{lower}(\phi_{A,B1})>-$ min-corr，则$(A,B2)$不是负相关的。

证明：由定理 1 可知，$\phi_{A,B1}\geqslant\text{lower}(\phi_{A,B1})$，$\phi_{A,B2}\geqslant\text{lower}(\phi_{A,B2})$。

由定理2可知，对相同的 $s(A)$，当 $s(B1)>s(B2)$，有 $\text{lower}(\phi_{A,B2})>\text{lower}(\phi_{A,B1})$。

如果 $\text{lower}(\phi_{A,B1})>-\text{min-corr}$，结合定理1和定理2可知：

$\phi_{A,B2}\geqslant\text{lower}(\phi_{A,B2})>\text{lower}(\phi_{A,B1})>-\text{min-corr}$，从而 $(A,B2)$不是负相关的。

当apriori性质不能用于相关项对的剪枝时，对含有 m 个项的事务，仅以 $m=1000$ 论，所有候选的2-相关项有 $C_m^2=m*(m-1)/2=499500$ 对，候选的3-相关项则有 $C_m^3=m*(m-1)*(m-2)/(3*2)=166167000$ 种组0合，…。由于受到内存空间的限制，统计和计算这么多组合的相关系数绝非易事。利用定理3我们可以减少候选负相关项对的个数。具体做法为：将 m 个项 $\{I_1, I_2, \cdots, I_m\}$ 按照每个项支持度的降序排列得到向量 $(Ia_1, Ia_2, \cdots, Ia_m)$，从 Ia_1 开始依次获得组合 $(Ia_1, Ia_2)\cdots(Ia_1, Ia_m)$；$(Ia_2, Ia_3)\cdots(Ia_1, Ia_m)$；…$(Ia_{m-1}, Ia_m)$。这样对每个以 $Ia_i(i=1,2,\cdots,m-1)$ 为首项的项对，如果 $\exists j(i+1\leqslant j\leqslant m)$，使得 $\text{lower}(\phi Ia_i, Ia_j)>-\text{min-corr}$，则根据定理3，任意的项对 $(Ia_i, Ia_{j+1},)$ 都不可能成为负相关对，可以从候选负相关项对的集合中排除。这样的一个 j 称为项 Ia_i 的停止点(stop point)，记为 st。

3.3 负相关规则的可靠度计算

若项对 (A,B) 之间存在负相关关系，则可能的规则形式有 $A=>\neg B$、$\neg A=>B$、$B=>\neg A$、$\neg B=>A$ 四种，因此需要根据可靠度来确定规则的前件与后件。如果直接按照可靠度的定义计算，则需要计算 $p(A)$，$p(B)$，$p(\neg A)$，$p(\neg B)$，$p(A,\neg B)$，$p(\neg A,B)$，$p(B,\neg A)$，$p(\neg B, A)$。即使考虑到 $p(\neg A)=1-p(A)$，$p(\neg B)=1-p(B)$，$p(A,\neg B)=p(\neg B, A)$，$p(\neg A,B)=p(B,\neg A)$，仍然需要计算 A、B、$(A,\neg B)$和$(\neg A,B)$的支持度。如前文所述，负项对的计算困难在于数据库中并没有保存顾客没有购买什么商品，事实上，如果单项 A 和 B 的支持度已知，则根据公式 $s(A,\neg B)=s(A)-s(A,B)$，$s(\neg A,B)=s(B)-s(A,B)$，这不仅解决了负项计数困难的问题，而且将 $(A,\neg B)$ 和 $(\neg A,B)$ 的两个支持度统一转换为求 (A,B) 的支持度。根据以上分析，我们推导可靠度计算公式如下：

$$r(A=>\neg B)=P(\neg B|A)-P(\neg B)=\frac{p(A,\neg B)}{p(A)}-p(\neg B)=\frac{s(A,\neg B)}{s(A)}-s(\neg B)=\frac{s(A)-s(A,B)}{s(A)}-(1-s(B))$$

同理，$r(\neg A=>B)=\dfrac{s(B)-s(A,B)}{1-s(A)}-s(B)$

$$r(B=>\neg A)=\frac{s(B)-s(A,B)}{s(B)}-(1-s(A))$$

$$r(\neg B=>A)=\frac{s(A)-s(A,B)}{1-s(B)}-s(A)$$

3.4 负相关规则挖掘算法

3.4.1 负相关项对挖掘算法 MNI

NegCand=ϕ; NCP=ϕ; // NegCand 为候选负项对集合，NCP 为负项对集合

scan database to get supports of all single items;

VEC=sorted items vector $(Ia_1, Ia_2, \cdots, Ia_m)$ according to descending support values of items;

for $i=1$ *to* $m-1$ do

$$\text{lower}(\phi, Ia_i, Ia_m)=-\frac{\sqrt{s(Ia_i)s(Ia_m)}}{\sqrt{(1-s(Ia_i))(1-s(Ia_m))}}$$

if (lower(ϕ, Ia_i, Ia_m)$\leqslant$ −min-corr) then

for $j=i+1$ to m do

NegCand= NegCand$\cup\{(Ia_i, Ia_j)\}$;

endfor

else

st=BinSearch(i, VEC); //binary search the stop point for Ia_i

for $j=i+1$ to st do

NegCand= NegCand$\cup\{(Ia_i, Ia_j)\}$;

endfor

endfor

scan database to get supports of all possible negative pairs (Ia_i, Ia_j) in NegCand;

for each (Ia_i, Ia_j) in NegCand do

$$\phi_{Ia_i, Ia_j}=\frac{s(Ia_i,Ia_j)-s(Ia_i)s(Ia_j)}{\sqrt{s(Ia_i)s(Ia_j)(1-s(Ia_i))(1-s(Ia_j))}}$$

if ($\phi_{Ia_i, Ia_j}\leqslant$ −min-corr)

then $NCP=NCP\cup\{(Ia_i, Ia_j)\}$;

endfor

return NCP

3.4.2 负相关规则挖掘算法 MNECR

$NCR=\phi$; //NCR 为负相关规则的集合

for each (Ia_i, Ia_j) in NCP

if ($r(Ia_i=>\neg Ia_j)\geqslant$ min-reli)

then $NCR=NCR\cup\{Ia_i=>\neg Ia_j\}$;

if ($r(\neg Ia_j=>Ia_i)\geqslant$ min-reli)

then $NCR=NCR\cup\{\neg Ia_j=>Ia_i\}$;

if ($r(Ia_j=>\neg Ia_i)\geqslant$ min-reli)

then $NCR=NCR\cup\{Ia_j=>\neg Ia_i\}$;

if ($r(\neg Ia_i=>Ia_j)\geqslant$ min-reli)

then NCR=NCR$\cup\{\neg Ia_i=>Ia_j\}$;

endfor

return NCR

4 算法 MNI 完全性及正确性证明

记全部负相关项对的集合为 $S1$，通过算法 MNI 获得的项对集合为 $S2$。

完全性：证明 $\forall x=(A, B)\in S1$，有 $x\in S2$。

证：采用反证法，假设 $\exists$ 某个负相关对 $(X,Y)\in S1$，且 $(X,Y)\notin S2$。

从算法 MNI 可以得知，对 $(Ia_i, Ia_j)\in$ NegCand，如果 $\phi_{Ia_i, Ia_j}\leqslant$ −min-corr，则有 $(Ia_i, Ia_j)\in S2$，由此可知 $S2\subseteq$ NegCand。若有上述假定中的负相关对 $(X,Y)\in S1$，且 $(X,Y)\notin S2$，那么此负相关对 $(X,Y)\notin$ NegCand。

如果 $(X,Y)\notin$ NegCand，则分析算法可以知道，此负相关对必然属于集合 $NN=\{(Ia_i, Ia_{g+1}),(Ia_i, Ia_{g+2}),\cdots,(Ia_i, Ia_m)\}(i=1,2,\cdots,m-1)$。但是根据定理3我们已经知道，集合 NN 中的任意一个项对都不是负相关的，这与假设矛盾。

正确性：证明 $\forall x=(A, B)\in S2$，有 $x\in S1$。

证：对 $\forall x=(A, B)\in S2$，根据算法必有 $\Phi_{A,B}\leqslant$ −min-corr，则按照负相关的定义，(A, B) 是负相关对，即 $x\in S1$。

5 实验结果及分析

若记 nItems 为项数，nBoundItems 为单项支持度为0或

1 的边界点个数，nCandPairs 为根据算法 MNI 裁剪后产生的候选负相关项对，则实际的修剪率可定义如下：

$$pruneRate=\frac{nCandpairs}{(nItems-nBoundItems)(nItems-nBoundItems-1)/2}$$

5.1 数据集及实验环境

由于使用综合数据产生程序生成的数据，其实际相关性非常小，本实验采用真实数据集。所使用的数据集中，Chess，Mushroom，Connect 数据集来自 IBM[9]，Pumsb，Pumsb Star 数据集来自 UCI Repository[10]。各数据集的特征如表 1 所示。本实验在 PentiumIV，RAM512M 机器上通过，OS 为 Windows2000。

表 1 数据集特征表

数据集 Data Set	项数 #Item	记录数 #Record
Chess	75	3196
Mushroom	119	8124
Connect	129	67557
Pumsb	2113	49046
Pumsb Star	2088	49046

5.2 裁剪率

从图 1 和图 2 可以看出，尽管各个数据集的裁剪率有所不同，但是随着最小相关系数阈值的增大，裁剪率在总体上都呈现出上升的趋势。

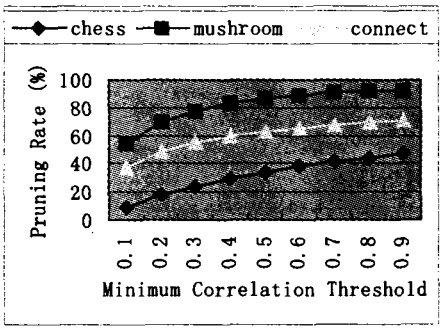

图 1 IBM 数据集裁剪效果图

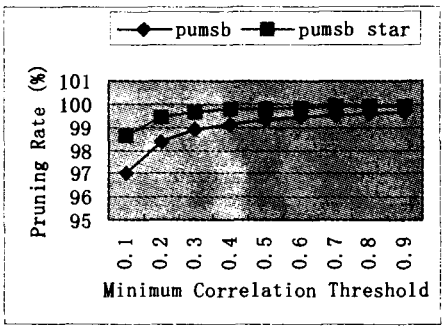

图 2 UCI 数据集裁剪效果图

5.3 运行时间

Connect，Pumsb，Pumsb Star 数据集上裁剪前后的运行时间比较如图 3、4、5 所示。裁剪前的运行时间除稍有波动外，基本表现为与 X 轴平行的直线，而裁剪后的运行时间随着最小相关度阈值的增大而呈现下降趋势。Chess 和 Mushroom 数据集上也表现出相同的特征，但是由于数据集太小，裁剪前后在运行时间上的改善不很明显，在此不再列出。

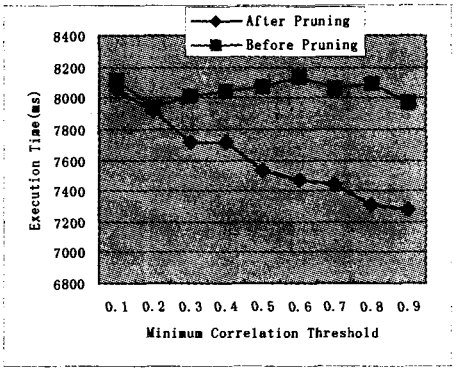

图 3 Connect 数据集运行时间图

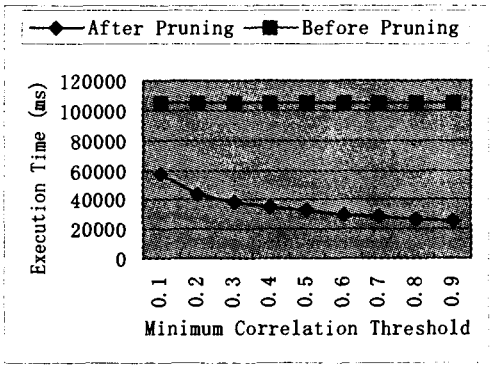

图 4 Pumsb 数据集运行时间图

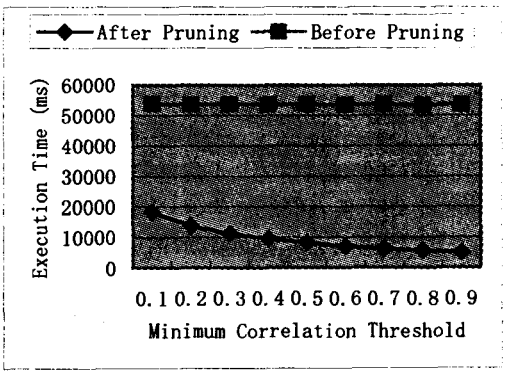

图 5 Pumsb Star 数据集运行时间图

5.4 负相关项对个数

在 IBM 和 UCI 数据集上产生的负相关项对个数如图 6 和 7 所示。

负相关项对的个数随着最小相关度阈值的增大而减少，这是可以预期的。比较有趣的是，我们可以发现当最小相关系数阈值设置到小于 0.3 后，挖掘到的负相关项对个数的增加趋势十分迅猛，而大于 0.3 后的负相关项对的个数的减少趋势则相对要平稳得多。这正如统计学所指出的，小于 0.3 的相关系数阈值太小，将导致许多相关性并不明显的项对入选。

结论及将来工作 本文提出的算法通过 phi 相关系数的下界来裁剪负相关项对的搜索空间。在真实数据集上的实验表明，裁剪率随着最小相关系数阈值的增大而增大，裁剪后的运行时间相对于裁剪前的运行时间大为缩短，而且这种效率上的提高随着最小相关系数阈值的增大而增大。与此同时，

（下转第 163 页）

1；banana 有 400 个训练样本，4900 个测试样本，2 维输入，输出维数是 1；Breast Cancer Data Set 有 200 个训练样本，77 个测试样本，输入维数是 9，输出维数是 1。这些数据都有 100 组训练样本和 100 组测试样本，这里任意选一组进行下面的实验。

首先采用文[7]中的二次规划下的方法进行实验，将该方法记为 QPMSVM。然后采用本文中的线性规划下的未分解算法（LPMSVM）和分解算法（LPDMSVM）进行对比实验。在分解算法中，将每类样本分成三组小样本集，然后对决策函数值采用加权平均，最后根据最大值判断新样本所属类别。

在 QPMSVM 方法中，取 $C=5,\sigma^2=0.3$。在本文的两种算法中采用相同的参数，其中，German 数据和 breast-cancer 数据中取 $C=5,\sigma^2=0.0001$，banana 数据中取 $C=5,\sigma^2=0.3$。三种算法的分类结果如表 3 所示。从表中可以看到，本文两种算法都保持了良好的分类精度，而且程序运行时间比二次规划下的算法运行时间大大缩短，尤其是分解算法运行速度更快。

表 3　实验结果

方法	German	banana	breast-cancer
QPMSVM	90% (125.5200)	87.29% (726.2550)	75.32% (16.4340)
LPMSVM	90% (78.2030)	88.86% (308.8140)	77.92%(7.2910)
LPDMSVM	90% (47.5080)	89.53% (248.7080)	77.92% (3.8850)

结论　本文根据一类分类思想，提出了一种基于线性规划的多类分类算法及其分解形式。对人工三螺旋线和实际数据的仿真实验结果表明该方法简单、运行速度快，而且仍然能保持良好的分类精度。当面对大规模数据分类问题时，可以采用分解算法来求解。另外，本文方法还定义了一个信任度函数，它可以在一定程度上对分类结果给出一个可信度评判，必要时可以根据信任度大小对分类结果进行适当调整。

参 考 文 献

1 Vapnik V N. Statistical Learning Theory [M]. New York, Wiley, 1998
2 Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition [R]. Knowledge Discovery and Data Mining, 1998, 2 (2): 121～167
3 Bennett K, Blue J. A support vector machine approach to decision trees [R]. Rensselaer Polytechnic Institute, Troy, NY: R. P. I Math Repot, 1997. 97～100
4 Platt J C, Cristianini N, Shawe-Taylor J. Large Margin DAGs for Multiclass Classification. In: Solla S A, Leen T K, Muller K R, eds. Advances in Neural Information Processing System 12 (NIPS 1999), Pittsburgh, PA, USA, Cambridge, MA, MIT Press, 1999. 547～553
5 李昆仑，黄厚宽，田盛丰. 一种基于有向无环图的多类 SVM 分类器. 模式识别与人工智能[J], 2003, 16(2): 164～168
6 Weston J, Watkins C. Multi-class Support Vector Machines [R]: [CSD-TR-98-04]. Royal Holloway University of London, 1998
7 孙德山，吴今培，肖健华. 一种新的多类分类算法. 模式识别与人工智能[J], 2004, 17(3): 357～361
8 Scholkopf B, Williamson R, Smola A, et al. Support Vector Method for Novelty Detection. http: // citeseer. nj. nec. com/ 400144. html
9 Ratsch G, Scholkopf B, Mika S, et al. SVM and Boosting: One Class. http: //citeseer. nj. nec. com/516656. html
10 Tax D. One-class classification: [PhD thesis]. Delft University of Technology, Netherlands, 2001
11 Mangasarian O L. Arbitrary-norm separating plane [J]. Operation Research Letters, 1999, 24(1): 15～23

（上接第 127 页）

所发现的负项对个数则随最小相关系数阈值的增大而减少。

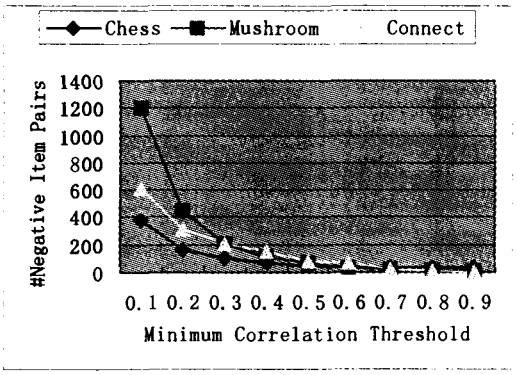

图 6　IBM 数据集负相关项对个数

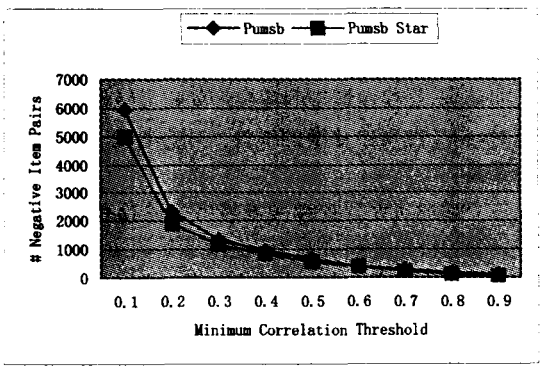

图 7　UCI 数据集负相关项对个数

下一阶段的研究将集中在算法的可扩展性和所发现的负规则的使用上。

参 考 文 献

1 Agrawal R, Srikant R. Fast algorithm for mining association rules in large databases. In: Proc. of 20th Int'l Conference on Very Large Databases (VLDB1994), 1994. 487～499
2 Han J, Pei J, Yin Y. Mining Frequent patterns without candidate generation. ACM SIGMOD, Dallas, Texas, 2000
3 Brin S, Motwani R, Silverstein C. Beyond Market Basket: Generalizing Association Rules to Correlation. In: Proc. 1997 ACM-SIGMOD Int'l Conf. Management of Data, 1998. 265～276
4 Savasere A, Omiecinski E, Navathe S. Mining for Strong Negative Associations in a Large Database of Customer Transactions. In: Proc. of the Fourteenth Intl. Conf. on Data Engineering (ICDE1998), 1998. 494 ～ 502
5 Ahmed K M, EI-Makky N M, Taha Y. A note on "Beyond Market Baskets: Generalizing Association Rules to Correlations". ACM SIGKDD Explorations, 2000, 1(2): 46～48
6 Wu X, Zhang C, Zhang S. Mining both positive and negative association rules. In: Proc. of 19th Intl. Conf. on Machine Learning (ICML2002), 2002. 658～665
7 Antonie M-L, Zaiane O R. An Associative Classifier based on Positive and Negative Rules. In: Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004
8 Xiong H, Shekhar S, Tan P-N, et al. Exploiting A Support-based Upper Bound of Phi's Correlation Coefficient for Identifying Strongly Correlated Pairs. In: Proc. of the tenth ACM SIGKDD conference, 2004
9 http://www. almaden. ibm. com/software/quest/Resources/ index. shtml
10 http://www. ics. uci. edu/～mlearn/MLRepository. html