基于抗体浓度和亲合度的关联规则挖掘算法

詹 芹,廖慧芬
ZHAN Qin,LIAO Hui-fen

九江学院 信息科学与技术学院,江西 九江 332005
School of Information Science and Technology,Jiujiang University,Jiujiang,Jiangxi 332005,China
E-mail:zhanqin2008@tom.com

ZHAN Qin,LIAO Hui-fen.Association rule algorithm based on concentration and affinity of antibodies.Computer Engineering and Applications,2009,45(21):147-149.

Abstract: Association rule is one of the important models of data mining,and has the most significant application value.This paper combining with the concentration and affinity of antibodies,brings forward a method of mining association rules based on clonal simulated annealing genetic algorithm.It first generates a new group of individuals through clonal operation,and makes mutation/selection independently with all the generated individuals respectively.Experiment results demonstrate that this method can solve association rule mining effectively.

Key words: data mining;association rule;concentration of antibodies;affinity

摘 要: 关联规则是数据挖掘的重要模式之一,有着极其重要的应用价值。基于抗体浓度和亲合度的选择策略,提出了一种克隆模拟退火遗传挖掘算法。该挖掘算法先通过克隆操作来产生一组新的抗体,然后再独立地对所产生的抗体进行变异和克隆选择操作,从而求得问题的最优解。实验结果表明该算法能高效地解决关联规则挖掘问题。

关键词: 数据挖掘;关联规则;抗体浓度;亲合度
DOI: 10.3778/j.issn.1002-8331.2009.21.043 **文章编号:** 1002-8331(2009)21-0147-03 **文献标识码:** A **中图分类号:** TP311

数据挖掘是从大量的无规律的繁杂数据中抽取出隐含的、具有潜在应用价值信息的过程。随着当今信息社会数据的爆炸式增长,人类分析数据和从中提取有用信息的能力远远不能满足实际需要。关联规则挖掘是解决这类问题的有效途径。关联规则的任务是发现大量数据中项集之间有趣的关联或相关联系[1],它引起了国内外众多学者的关注,成为目前数据挖掘研究领域的一个焦点。

关联规则是数据项之间存在的规则,是在同一事件中出现的不同项之间的相关性。设数据项集 $I=\{i_1,i_2,\cdots,i_n\}$,事务数据集 $T=\{T_1,T_2,\cdots,T_m\}$,其中 $T_k(1\leqslant k\leqslant m)$是事务数据集 T 的数据项,也是数据项集 I 中的数据项,并且 $T\subseteq I$ 的一种关联规则是形如 $X\Rightarrow Y$ 的蕴涵关系,其中 $X\subset I,Y\subset I$,并且 $X\cap Y=\varnothing$。如果在 T 中,包含 $X\cup Y$ 事务所占比例为 $S\%$,则称 $X\Rightarrow Y$ 有支持度 $Supp(X\Rightarrow Y)=Supp(X\cup Y)=P(X\cup Y)$。如果在 T 中,$C\%$的事务包含 X 同时也包含 Y,则称 $X\Rightarrow Y$ 有可信度 $Conf(X\Rightarrow Y)=P(Y|X)=Supp(X\cup Y)/Supp(X)$。支持度则是对关联规则的重要性的度量,而可信度是对关联规则的准确度的度量。一般地,用户可以定义两个阈值,分别为最小支持度阈值和最小可信度阈值。当挖掘出来的关联规则的支持度和可信度都满足这两个阈值,就认为这条规则是有效的,否则是无效的。

1 现有的关联规则挖掘算法

目前,围绕着数据挖掘中的关联规则挖掘方法,国内外已经做了大量的研究工作,先后提出各种启发式挖掘算法,其中比较流行的有遗传算法[2-4](Genetic Algorithm,GA)和模拟退火遗传算法[5](Simulated Annealing Genetic Algorithm,SAGA)等。下面具体分析这两种挖掘算法。

1.1 GA 挖掘算法

GA 是模拟生物遗传和进化过程建立起来的一种随机搜索和优化算法。GA 挖掘算法是一种在挖掘的过程中不产生频繁集候选项的关联规则挖掘方法,在数据挖掘中应用得非常普遍。文献[2]中,GA 挖掘算法采用实数数组的方法进行编码,实数数组的元素个数与事务数据库中的字段的个数相对应,元素值代表字段的属性值。该算法利用精英重组,一致变异以及自适应参数的手段进行数据挖掘。虽然 GA 挖掘算法全局寻优能力强,在实际应用过程中取得了比较好的实验结果,但是它容易陷入局部最优解。

1.2 SAGA 挖掘算法

SAGA 算法是将 GA 与 SA(Simulated Algorithm,SA)算法结合构成的一种混合优化算法。文献[4]中 SAGA 挖掘算法采用实数串编码方式,每条染色体由一组实数构成,每个实数都与事务数据库中的一个字段值相对应。在进化过程中,采用自适

作者简介: 詹芹(1977-),女,讲师,主要研究领域为数据挖掘;廖慧芬(1977-),女,讲师,主要研究领域为人工智能。
收稿日期: 2009-05-04 **修回日期:** 2009-06-22

应方式动态选取交叉和变异概率,并将适应度函数描述为:

$$fitness(x)=W_s\times\frac{Supp(x)}{Supp_{\min}}+W_c\times\frac{Conf(x)}{Conf_{\min}}$$

其中,$W_s+W_c=1$,$W_s\geqslant 0$,$W_c\geqslant 0$,$Supp_{\min}$是支持度的阀值,$Conf_{\min}$是可信度的阀值。

SAGA 算法虽然能有效抑制 GA 的早熟收敛现象,但是该算法的选择操作只能选取优胜个体,而不能充分克隆优秀个体,同时动态选取交叉和变异概率,使算法实现起来比较困难。

在文献[2-5]研究成果的启发下,融入克隆思想,并首次将 GA,SA 和克隆选择算法[6-7](Clonal Selection Algorithm,CSA)的优点进行结合,提出了一种基于克隆模拟退火遗传策略(Clonal Simulated Annealing Genetic Algorithm,CSAGA)的挖掘算法。实验结果显示该方法能高效地解决关联规则挖掘问题。

2 CSAGA 挖掘算法

2.1 CSAGA 算法结构

在 CSAGA 挖掘算法中,采用文献[2]GA 挖掘算法的实数数组编码方法。实数数组的元素个数与事务数据库中的字段的个数相对应,元素值代表了字段的属性值。用一个长度为 n 的数组来表示事务数据库的抗体编码,$a[1]$表示字段 1,$a[2]$表示字段 2,…,$a[n]$表字段 n,其中 0 值表示此属性与其他的属性无关联。因此,CSAGA 算法中的抗体可以描述为:$a[1]a[2]\cdots a[n]$。

亲和度函数是用来表明抗体与抗原之间的匹配程度,亲和度越高,说明抗体越接近所求问题的解。借鉴文献[5]SAGA 挖掘算法的适应度函数,CSAGA 挖掘算法的亲和度函数定义如下:

定义 1(亲和度函数)

$$affinity(x)=W_s\times\frac{Supp(x)}{Supp_{\min}}+W_c\times\frac{Conf(x)}{Conf_{\min}}$$

其中,$W_s+W_c=1$,$W_s\geqslant 0$,$W_c\geqslant 0$,$Supp_{\min}$是支持度的阀值,$Conf_{\min}$是可信度的阀值。

在 CSAGA 挖掘算法中,克隆算子能有效扩大群体的规模。每个抗体与抗原的亲和度越大,抗体的克隆规模也就越大。抗体群 A 中每一个抗体 a_i 按规模 N_c 克隆得到新的抗体群,$N_c=\alpha\times f(a_i)/\sum_{i=1}^{n}f(a_i)$,$i=1,2,\cdots,n$,$\alpha$ 是放大系数。

通过克隆扩大了群体的规模后,对克隆后的抗体群中每个抗体进行变异,可以提高群体中抗体的多样性,扩大搜索范围,用来寻找更优秀的抗体。采用文献[2]GA 挖掘算法的一致变异算法。假如 a_i 和 a_i'分别代表变异前的父抗体和子抗体,然后根据文献[4]SAGA 挖掘算法的个体模拟退火接受规则,若 $\min\{1,\exp(-(f(a_i')-f(a_i))/T_k)\}>P_m$,则接受新的抗体;否则,放弃变异后的抗体 a_i'。T_k 为第 k 次进化的温度,P_m 为变异概率。$f(a_i)$,$f(a_i')$分别是抗体 a_i 和 a_i'的亲和度。

克隆选择操作是从经变异后的各自子代和相应父代中选择优秀的抗体,从而形成新的种群。利用抗体的浓度和亲和度来进行克隆选择操作。假设抗体群中抗体 a_i 的数目为 r 个,则抗体 a_i 的浓度和选择概率分别定义如下:

定义 2(抗体 a_i 浓度)

$$\eta=\frac{r}{M}$$

定义 3(抗体 a_i 的选择概率)

$$p_s(i)=\lambda\times\frac{affinity(a_i)}{\sum_{i=1}^{M}affinity(a_i)}+(1-\lambda)\times\frac{r}{M}$$,其中,$p_s(i)$和 $affinity(a_i)$分别是抗体 a_i 的选择概率和亲和度函数,M 为抗体规模,λ 为克隆选择系数,$0.5\leqslant\lambda<1$。然后采取“轮盘赌”式的选择策略:

(1)计算所有抗体的选择概率;

(2)随机生成一个数 $r=random[0,1]$;

(3)若 $p_s(1)+\cdots+p_s(i-1)<r<p_s(1)+\cdots+p_s(i)$,则第 i 个抗体被选择到下一代抗体群。

这种克隆选择策略可以有效选择进化亲合度大的抗体,同时抑制浓度大的抗体,保证了进化群体中抗体的多样性,避免早熟收敛问题。

2.2 CSAGA 算法流程

CSAGA 挖掘算法流程如下:

(1)根据实验中数据库提供的信息,产生初始抗体群 $A=\{a_1,a_2,\cdots,a_N\}$,并进行编码;同时初始化参数:种群规模 M,变异概率 P_m,初始温度 T_0,进化代数 $k=0$,W_s,W_c,α,λ。

(2)计算抗体的可信度,支持度和亲和度。

(3)对抗体群分别进行克隆,变异和克隆选择操作。

(4)$k=k+1$;当 $T_k\approx 0$ 时,算法结束,并提取关联规则;否则,$T_{k+1}=T_k\times(1-k/M)$,算法返回到步骤(2)。

3 CSAGA 算法的可行性分析

定理 1 CSAGA 算法的种群序列$\{X_k,k\geqslant 0\}$,是有限齐次马尔可夫链。

证明 1990 年 Eiben 提出 GA 的一种抽象表示,将进化过程定义为马尔可夫链,利用转移概率矩阵相乘来进行状态的变换,分析得出 GA 收敛到最优解的条件。CSAGA 与 GA 类似,假设染色体长度为 N,种群规模为 M。对于染色体的取值是离散的 0、1 的 GA,种群所在的状态空间大小是 2^{MN},由于初始化种群 $A(0)$的取值是连续的,理论上 CSAGA 中种群所在的状态空间是无限的,但 $A(0)$是有限精度的,设其维数为 v,则种群所在的状态空间大小为 v^{MN},因此种群是有限的,而算法中采用的克隆,变异,交叉都与 k 无关,所以 X_{k+1} 仅与 X_k 有关,即$\{X_k,k\geqslant 0\}$是有限齐次马尔可夫链。

定理 2 对于 CSAGA 算法马尔可夫链序列的种群满意值序列是单调不减的,即对于任意的 $k\geqslant 0$,有 $f(X_{k+1})\geqslant f(X_k)$,种群中的任何抗体都不会退化。

证明 在 CSAGA 算法中,采用保留最优个体来进行克隆操作,因此保证了每一代个体都不会退化。

定理 3 CSAGA 算法是以概率 1 收敛的。

证明 将 CSAGA 算法的状态转移用马尔可夫链来描述,并且将规模为 M 的群体认为是状态空间 S 中的某个点,用 $s_i\in S$ 表示 s_i 是 S 中的第 i 个状态,$s_i=\{x_1,x_2,\cdots,x_n\}$,显然 X_k^i 表示在第 k 代时种群 X_k 处于状态 s_i,其中随机过程$\{X_k\}$的转移概率为 $p_{ij}(k)$,则 $p_{ij}(k)=p\{X_{k+1}^j/X_k^i\}$,$I=\{i|s_i\cap s^*\neq\varnothing\}$。根据定理 2 可以得出:$f(X_{k+1}^j)\geqslant f(X_k^i)$,所以 $p_{ij}(k)>0$。设 $p_i(k)$为种群 X_k 处在状态 s_i 的概率,$p_k=\sum_{i\notin I}P_i(k)$,则由马尔可夫链的性质可知:

$$p_{k+1}=\sum_{s_i\notin S}\sum_{j\notin I}P_i(k)P_{ij}(k)=\sum_{i\in I}\sum_{j\notin I}P_i(k)P_{ij}(k)+\sum_{i\notin I}\sum_{j\notin I}P_i(k)P_{ij}(k),$$

由于 $\sum_{i\notin I}\sum_{j\in I}P_i(k)P_{ij}(k)+\sum_{i\notin I}\sum_{j\notin I}P_i(k)P_{ij}(k)=\sum_{i\notin I}p_i(k)=p_k$,因此,

$\sum_{i\notin I}\sum_{j\notin I}P_i(k)P_{ij}(k)=p_k-\sum_{i\notin I}\sum_{j\in I}P_i(k)P_{ij}(k)$。

根据以上公式,可以得出 $0\leqslant p_{k+1}<\sum_{i\in I}\sum_{j\notin I}P_i(k)P_{ij}(k)+p_k=p_k$,因此,$\lim_{k\to\infty}p_k=0$。又因为$\lim_{k\to\infty}p\{f_k=f^*\}=1-\lim_{k\to\infty}\sum_{i\notin I}p_i(k)=1-\lim_{k\to\infty}p_k$,可知$\lim_{k\to\infty}p\{f_k=f^*\}=1$。即所有包含在全局最优状态中的概率收敛为 1。

通过定理 1~3 的推导证明,表明 CSAGA 算法是完全可行的。

4　实验及实验参数配置

为了验证提出的观点和方法的有效性和可扩展性,在 Pentium® 4 CPU 2.80 GHz and 2 G 内存的 PC 机上,对文献[2] GA 挖掘算法,文献[5]SAGA 挖掘算法和本文 CSAGA 挖掘算法做了对比实验。利用某医院的病例数据库,进行相应的关联规则挖掘。通过运用 GA,SAGA 和 CSAGA 三种挖掘算法发现某种病情都与哪些因素密切相关,如冠心病易发生于中老年男性人群等规则。

实验过程中用到的一些参数配置如下:种群规模 150,交叉概率 0.85,变异概率 0.1,初始温度为 1,支持度阈值为 10%,可信度阈值为 70%,W_s=0.5,W_c=0.5,α=100,λ=0.8。

图 1~2 分别显示的是三种挖掘算法产生的关联规则数目和规则准确率的比较。假设在产生的 n 条关联规则中,其中与实际情况相吻合的规则数目为 m,则规则准确率为 m/n。如规则(乳腺癌⇒中年人,男),它表示的含义是中年男性容易患乳腺癌,显示这条规则是错误,而规则(乳腺癌⇒中年人,女)是符合实际情况的。

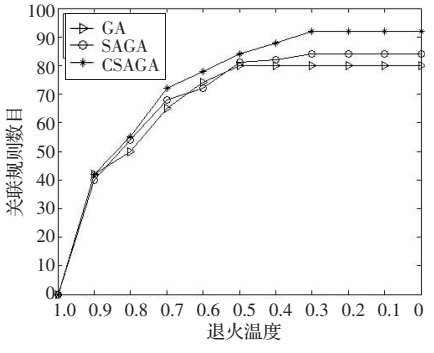

图 1　三种算法挖掘的关联规则数目

从图 1 中可以看出,CSAGA 挖掘算法比 GA 和 SAGA 挖掘算法更适合关联规则挖掘,因为采用抗体浓度和亲合度选择策略,促使进化过程中的优秀抗体保留到下一代抗体群,表 1 中列举的部分规则与实际情况相吻合,说明 CSAGA 算法挖掘

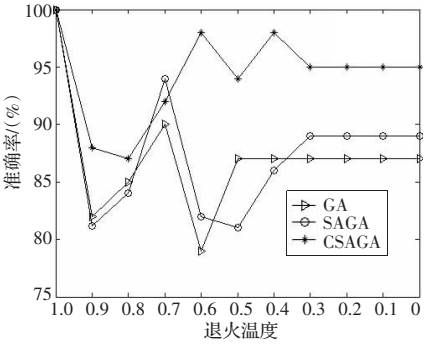

图 2　三种算法挖掘产生的规则准确率

关联规则是有效可行的。图 2 充分表明 CSAGA 挖掘算法的准确率较高(CSAGA 的准确率平均达到 92.5%,而 SAGA 的平均准确率只有 89.2%,GA 的平均准确率才 87.2%),这是因为 CSAGA 算法的克隆操作将解空间中的一个点分裂成多个相同的点,从而扩大解空间搜索范围,有利于在更大范围内搜索最优解。

5　结论

在关联规则挖掘中引入抗体浓度和亲合度的选择策略,提出了一种克隆模拟退火遗传挖掘算法,通过克隆,变异和克隆选择操作,发现隐藏在数据库中的规律。理论分析和仿真实验表明该算法能有效、快速地解决关联规则挖掘问题。在今后的研究工作过程中,将更好地研究和应用 CSAGA 挖掘算法到不同领域,帮助人们更好的发现和识别各种潜在的规律。

参考文献:

[1] 贾彩燕,陆汝钤.关联规则挖掘的取样误差量化模型和快速估计算法[J].计算机学报,2006,29(4):625-634.

[2] 赵方方,刘万军,陈芳元.遗传算法在关联规则挖掘中的应用研究[J].沈阳理工大学学报,2006,25(4):51-54.

[3] 王礼刚,左源瑞,李盛瑜.一种基于改进型遗传算法的关联规则提取算法及其应用[J].重庆师范大学学报:自然科学版,2006,23(2):42-45.

[4] Riyaz S,Selwyn P.Efficient genetic algorithm based data mining using feature selection with hausdorff distance[J].Information Technology and Management,2005,6(4):315-331.

[5] 武兆慧,张桂娟,刘希玉.基于模拟退火遗传算法的关联规则挖掘[J].计算机应用,2005,25(5):1009-1011.

[6] 赵春晖,孙莉.基于克隆选择算法的层叠滤波器的优化设计[J].哈尔滨工程大学学报,2007,28(4):456-460.

[7] Castro L N,Von Zuben F J.Learning and optimization using the clone selection principal[J].IEEE Transaction Evolution Computing,2002,6(3):239-251.