基于半空间和GA的关联规则快速挖掘算法

方德洲,李 淼
FANG De-zhou,LI Miao

1.中国科学院 合肥智能机械研究所,合肥 230031
2.中国科学技术大学 信息科学技术学院,合肥 230027
1.Institute of Intelligent Machines,Academia Sinica,Hefei 230031,China
2.School of Information Science and Technology,University of Science and Technology of China,Hefei 230027,China
E-mail:fangfdz@mail.ustc.edu.cn

FANG De-zhou,LI Miao.Fast algorithm for mining association rules based on half-spaces and GA.Computer Engineering and Applications,2007,43(2):193-195.

Abstract: This paper has proposed a fast algorithm for mining association rules based on half-spaces and GA.Most traditional algorithms are constrained by many aspects of practice,such as data type,real-life significance and so on,which probably constrain the capacity of knowledge acquisition seriously.By contraries,the algorithm proposed can get rid of the constraints.Moreover, besides supplying some interesting rules to user,it also does well in mining for large data set.

Key words: half-spaces;GA;association rules;data mining

摘 要:提出了一种利用半空间模型和遗传算法(GA)对关联规则进行快速挖掘的方法。传统关联规则挖掘算法往往受到数据类型、关联规则的实际意义等约束,大大限制了知识获取的能力。而此方法不再受到上述限制的困扰,并且可以挖掘出用户感兴趣的规则,尤其对于大规模样本集的效果也是相当不错的。

关键词:半空间;遗传算法;关联规则;数据挖掘
文章编号:1002-8331(2007)02-0193-03 **文献标识码**:A **中图分类号**:TP391

1 引言

关联规则主要描述大量数据中数据项间的相互关系,所以它作为一种潜在的知识而被广泛地应用于各个领域。因此,如何从大量数据中挖掘到关联规则及如何快速准确地获得用户所感兴趣的规则已经成为研究的重点。

随着神经网络、模糊算法和演化算法为代表的智能计算方法的引入,大大拓宽和丰富了关联规则的研究领域[2]。本文提出的算法是将半空间模型和遗传算法 (Genetic Algorithm,简称GA)相结合而产生的HSGA算法(Half-Space Genetic Algorithm),首先利用半空间模型构造出适应度函数,然后利用遗传算法对空间超平面进行全局搜索, 从而完成规则挖掘过程。将二者的优势互补,获得更加出色的挖掘效果。

2 半空间模型及其工作原理

所谓半空间模型就是指利用一个不封闭的超平面 $\alpha:f(\boldsymbol{x})=\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{x}-\alpha_0,=0$ 其中,$\boldsymbol{\alpha}=(\alpha_1,\cdots,\alpha_n)^{\mathrm{T}},\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}},n$ 为样本的属性个数,它将样本空间分成两个子空间,每个样本只能分布在其中一个子空间里。

2.1 半空间模型简介

半空间模型从空间上将所有的样本分成两类,对于任意一个样本 $\boldsymbol{x}_i\in X$(X 是一个含有 m 个样本的样本集合,$i=1,2,\cdots,m$)都会有两种情况:$f(\boldsymbol{x}_i)\geqslant 0$ 或 $f(\boldsymbol{x}_i)<0$,由此可以推广至多个超平面 $f(\boldsymbol{x}_i)=0(i=1,2,\cdots)$ 分割样本空间的情形,当这多个样本空间的相互位置适当时,可能使大多数样本同时满足某一种情况,例如 $f_1(\boldsymbol{x})\geqslant 0,f_2(\boldsymbol{x})<0,f_3(\boldsymbol{x})<0,\cdots$,而这恰恰满足了关联规则描述"如果发生了,则 B 也很可能发生"。由此可见,这种半空间模型方法非常适用于连续变量,因为它能够很轻易地用代数表达式将各个连续变量间的关系表示出来,而不需要将它们进行离散化处理。

2.2 工作原理

在实际生活中会有这样的情况,当对风速、温度、日照强度和寒冷指数(分别用表示)进行规则挖掘时,由于前三个要素都是影响寒冷指数的重要指标, 因此必须对它们进行综合考虑。而这种半空间的模型正好适用于这种规则的挖掘,可能挖掘出这样的规则:

"$x_1-0.2x_2-1.5x_3\geqslant 5\Rightarrow x_4>3$"

这表示了当风速,温度,日照强度满足左边的不等式时,寒冷指数很可能大于3。

对"$f_1(\boldsymbol{x})\geqslant 0\Rightarrow f_2(\boldsymbol{x})\geqslant 0$"这样的规则进行原理阐述。由这样的规则形式可知, 需要用两个超平面 α 和 β 对样本空间进

基金项目:国家863高技术研究发展计划资助项目(2003AA118040)。
作者简介:方德洲(1981-),男,硕士研究生,研究方向为模式识别与智能系统;李淼(1955-),女,研究员,博士生导师,研究方向为人工智能与农业知识工程。

行划分。但很可能得到这样的规则"$\alpha_1x_1+\alpha_2x_2\geqslant\alpha\Rightarrow\beta_1x_1+\beta_2x_2\geqslant\beta_0$",在大多数情况下并不需要这样的规则。这是由于 α 和 β 的高度相关所导致的[5],因此要想获得理想的关联规则,α 和 β 的位置关系相当重要,它们应该是相互垂直的,即 $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\beta}=0$。

当样本点落入 α 右边 β 上边时,都会增加关联规则的可信度;当落入其它区域时,都会降低可信度(如图 1)。但是如果可信度过高会意味着绝大多数样本都满足规则,这类规则反映的是众所周知的结论,而这样的规则往往不是人们感兴趣的[5]。

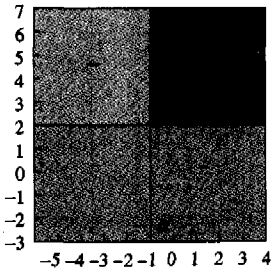

图 1 样本分布对规则可信度的影响

另一个问题就是支持度的选择,如果规则的支持度过高,会使得规则反映的事实可能是非常独立于样本集的;反之,规则可能只是反映了样本集中极个别的现象[4]。

因此,如何选择超平面 α 和 β 的位置来控制可信度和支持度,从而获得用户感兴趣的关联规则成为关键的问题。解决这类问题可以用全局搜索算法——GA 来解决。

3 遗传算法设计方案

遗传算法(GA)是建立在自然选择和自然遗传学机理基础上的迭代自适应概率性搜索算法。GA 具有不需要求梯度、能得到全局最优解、算法简单、可并行处理等优点,GA 已成功应用于各种复杂问题的优化中,在许多传统优化技术难以解决的场合更显示出其优越性。本文就是利用 GA 的这些优点来寻求超平面 α 和 β 的最优位置。

3.1 个体编码

所期望的关联规则是有实际意义的,因此将所关心的并且组合起来有意义的属性放在一起,这样就构成了一个兴趣属性集 I,例如对于一个有 n 个属性样本的数据集,它的兴趣属性集可能表示为:$I=\{\{x_1,x_2,x_3\},\{x_4,x_6\},\{x_5,x_{n-1},x_n\}\}$,注意到的任意两个子集的交集都是 Φ,这些子集都是用户所感兴趣的属性集合。

每个个体都是由超平面 α 和 β 构成,每个超平面的编码长度有 $m+n$ 位且都采用 0/1 编码方式,前 m 位是该超平面的常数项,初始值取原点到它的距离,其中第一位是符号位,因此这个距离是有符号的距离。m 为所有样本到原点的最大距离的二进制位数。后面的 n 位对应为样本 n 个属性的系数。初始化时,这两个超平面分别从兴趣属性集 I 中选取一个不同的子集,并将相应位置 1,其余位置 0,这样使得 $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\beta}=0$,从而保证了 α 和 β 的相互垂直关系。例如 β 选取 β_{n-1} 和 β_n 为 1,则 α_{n-1} 和 α_n 一定为 0(如图 2)。

由此可见,个体编码就是对超平面 α 和 β 的各个系数进行编码,其总长度有 $2(m+n)$ 位。

3.2 适应度函数

适应度函数是评估个体优劣的标准,它决定了最终遗传结果的性能[1]。本文提出的适应度函数是根据所有样本与超平面

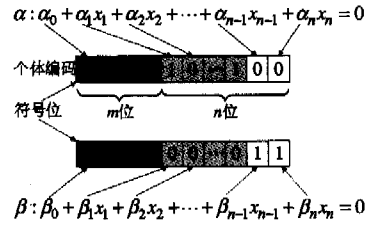

图 2 超平面 α 和 β 的编码规则

α 和 β 之间的位置关系来对个体进行评估的。

对整个样本集合,超平面 α 和 β 构成的关联规则的可信度函数为:

$$l(\boldsymbol{\alpha},\boldsymbol{\beta},X)=\frac{1}{2n}\sum_{x\in X}^{n}(\sigma(\delta(\boldsymbol{\alpha},\boldsymbol{x}))+(\sigma(\delta(\boldsymbol{\beta},\boldsymbol{x})))$$

它描述了样本点合落入"+"区域(如图 1)的情况。其中,$\sigma(x)$是改进的 *sigmoid* 函数,d 为所有样本到原点的最大距离,$\delta(\boldsymbol{\alpha},\boldsymbol{x})$是样本到超平面 α 的有符号距离,它们分别表示为:

$$\sigma(x)=\frac{1}{1+e^{-\frac{x}{\sqrt{d}}}}\quad x\in R$$

$$\delta(\boldsymbol{\alpha},\boldsymbol{x})=\frac{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\beta}}{|\boldsymbol{\alpha}|}-\alpha_0$$

超平面 α 和 β 构成的关联规则的支持度函数为:

$$c(\boldsymbol{\alpha},X)=\sigma\left(\left|\sum_{x\in X}\mathrm{sgn}(\delta(\boldsymbol{\alpha},\boldsymbol{x}))\right|\right)$$

它描述了样本点合落入超平面 α 右边的情况。其中,$\mathrm{sgn}(x)$是符号函数。

超平面 α 和 β 构成的关联规则的对比度函数为:

$$r(\boldsymbol{\alpha},\boldsymbol{\beta},X)=\sigma\left(\left|\sum_{x\in X}I(\delta(\boldsymbol{\alpha},\boldsymbol{x})>0)\mathrm{sgn}(-\delta(\boldsymbol{\beta},\boldsymbol{x}))\right|\right)$$

它描述了样本点集合落入超平面 β 下方的情况。因为不能将 β 无限地向下移动,使所有的样本点都分布在 β 的上方,由于这样的规则是没有用的,因此用这个函数来限制超平面 β 的位置。其中,$I(x)$表示当 x 为真时,返回 1;否则为 0。

超平面 α 和 β 构成的关联规则的补偿函数为:

$$e(\boldsymbol{\alpha},\boldsymbol{\beta},X)=\frac{1}{2}\left(\sum_{\gamma=\alpha,\beta}\exp\left(-\frac{\left|\gamma_0-\sum_{x\in X}\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{x}/m\right|}{\max(\boldsymbol{\gamma}^{\mathrm{T}}X)}\right)\right)$$

其中,m 是样本个数。通过支持度函数和对比度函数来选取 α_0 和 β_0 还是不够的,用个体和所有样本进行比较得到个体与样本的差异。差异越大,个体越差;反之,个体越好。对于较好的个体,利用这个激励函数来提高它的适应度。

由上述四个标准所定义函数的线性组合构成了本算法的适应度函数:

$$L(\boldsymbol{\alpha},\boldsymbol{\beta},X)=\lambda_1\cdot l(\boldsymbol{\alpha},\boldsymbol{\beta},X)+\lambda_2\cdot c(\boldsymbol{\alpha},X)+\lambda_3\cdot r(\boldsymbol{\alpha},\boldsymbol{\beta},X)+\lambda_4\cdot e(\boldsymbol{\alpha},\boldsymbol{\beta},X)$$

3.3 遗传算子

本文旨在阐述半空间模型在关联规则挖掘上的应用,所以使用了标准 GA 与之相结合。选择算子采用的是赌轮选择方法。交叉算子采用了两点交叉方式,即对每个超平面的常数项和兴趣属性进行交叉操作。变异算子分别对每个超平面的常数项和兴趣属性的某一位进行变异。为了防止早熟情况的发生,本算法将在每一代中增加新的样本,从而保证了个体的多样性。详细方法请阅读文献[1]。

4 实验结果

为了评估本文提出的基于半空间和 GA 的关联规则挖掘算法的性能,对该方法进行了仿真实验。为了与国家"863"计划资助项目紧密结合,本文选取的实验样本来源于云南水稻数据库,它记录了近 50 多年来云南气候、土壤、水稻、海拔等各种综合信息。在使用它之前要进行规格化处理。

各个遗传参数的选择方式如下:

种群规模,遗传代数分别为:50,100

选择概率,交叉概率分别为:0.6,0.35

变异概率概率:0.05

适应度函数的系数分别为:$\lambda_1=0.4,\lambda_2=0.2,\lambda_3=0.2,\lambda_4=0.2$

从图 3 中可以看出,HSGA 的性能要明显比标准的 GA[3] 好。HSGA 具有收敛速度快、适应度高的优点,并且它的平均适应度也比较高,这就意味着 HSGA 一次能够搜索到多个用户感兴趣的关联规则,且它们有着很好的实际意义。

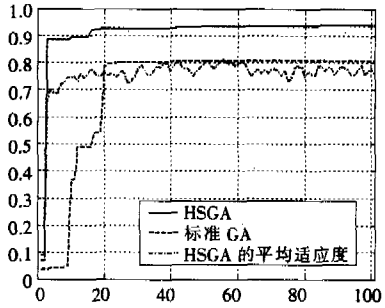

图 3 HSGA 和标准 GA 的适应度

HSGA 与标准 GA[3]在性能上的比较如表 1 所示。采用的是拥有 581 012 个样本的集合,它来源于 UCI 数据库[6]。选择其中 10 个属性来进行挖掘,整个过程运行于 Pentium IV 2.9 GHz 机器之上。从表 1 的结果来看 HSGA 能够以较短的时间在样本集中搜索到更多的关联规则,同时这些规则的可信度也比标准 GA 挖掘的规则要高。而基于标准 GA 的挖掘算法在样本集规模扩大后的性能急剧下降。

表 1 HSGA 与标准 GA 的性能对比

样本规模/千条	HSGA			GA		
	规则个数	平均可信度	时间/s	规则个数	平均可信度	时间/s
0.1	23	0.910	2.8	21	0.820	3.7
1.0	23	0.871	9.4	19	0.792	13.0
5.0	21	0.836	26.0	15	0.754	36.0
10.0	19	0.831	68.0	12	0.736	87.0
50.0	20	0.813	197.0	13	0.701	385.0
100.0	19	0.806	524.0	9	0.656	986.0

5 结束语

本文在关联规则挖掘领域提出了一种新的方法——HSGA 算法,它将半空间模型与遗传算法相结合。该算法的优点在于:

(1)它能够在多个相关属性中提取关联规则,而不需要考虑各个属性间的相互关系。

(2)它无需对数据进行离散化处理,因此这种算法特别适用于连续型数据。

(3)HSGA 使用了选择兴趣属性集的策略,这种启发式方法使得挖掘出的规则不再盲目,而是能够满足用户的期望。

HSGA 算法充分发挥了半空间模型和 GA 的优点,使其在关联规则挖掘领域表现得更为出色。通过实验证明了该算法是可行有效的。(收稿日期:2006 年 8 月)

参考文献:

[1] 陈国良,王煦法,庄镇泉,等.遗传算法及其应用[M].北京:人民邮电出版社,1996:28-97.

[2] 朱明.数据挖掘[M].合肥:中国科学技术大学出版社,2002:5-30.

[3] 彭建.一种基于遗传算法的关联规则挖掘方法[J].计算机技术与自动化,2005,24(2):75-77.

[4] Aumann Y,Lindell Y.A statistical theory for quantitative association rules[J].Journal of Intelligent Information Systems,2003,20(3):255-283.

[5] Ruckert U,Richter L,Kramer S.Quantitative association rules based on half-spaces:an optimization approach[C]//Proceedings of the Fourth IEEE International Conference on Data Mining,2004:507-510.

[6] Blake C,Merz C.UCI repository of machine learning databases,1998.

(上接 98 页)

及其父节点的方式,节省大量的比较时间,实验表明,本文所提出基于同义概念的概念格纵向合并算法是有效的。并且,同域概念格的纵向合并算法也很适用于对概念格进行并行构造,将本算法真正地用于概念格分布并行构造是将要进行的下一步工作。(收稿日期:2006 年 5 月)

参考文献:

[1] Ganter B,Wille R.Formal concept analysis:mathematical foundations[M].Berlin:Springer-Verlag,1999.

[2] Krohn U,Davies N J,Weeks R.Concept lattices for knowledge management[J].BT Technol J,1999,17(4):108-113.

[3] Kuznetsov S O.Machine learning on the basis of formal concept analysis[J].Automation and Remote Control,2001,62(10):1543-1564.

[4] 沈夏炯,刘宗田.形式概念分析方法与软件过程改进[J].计算机科学,2003(7):103-105.

[5] Godin R,Missaoui R,Alaoui H.Incremental concept formation algorithms based on Galois(concept)lattices[J].Computational Intelligence,1995,11(2):246-267.

[6] 沈夏炯,韩道军,刘宗田,等.概念格构造算法的改进[J].计算机工程与应用,2004,40(24):100-103.

[7] 李云,刘宗田,陈峻,等.基于属性的概念格渐进式生成算法[J].小型微型计算机系统,2004,25(10):1768-1771.

[8] 张凯,胡运发,王瑜.基于互关联后继树的概念格构造算法[J].计算机研究与发展,2004,41(9):1493-1499.

[9] Li Yun,Liu Zong-tian.Theoretical research on the distributed construction of concept lattices [C]//Proceedings of the Second International Conference on Machine Learning and Cybernetics.Xian:Institude of Electrical and Electronics,2003:474-479.

[10] Liu Zong-tian,Li Liang-sheng,Zhang Qing.Research on a union algorithm of multiple concept lattices[C]//RSFDGrC 2003,LNAI 2639.Berlin:Springer-Verlag Heidelberg,2003.533-540.

[11] 李云,刘宗田,陈峻,等.多概念格的横向合并算法[J].电子学报,2004,11(11):1849-1854.