基于模糊关联规则的预测系统遗传优化*

李 燕,王 锋
(杭州电子科技大学 计算机学院,浙江 杭州 310018)

摘要:为提高预测系统中的预测精度,提出了一种基于模糊关联规则的优化的预测系统设计方法。该方法通过两个阶段来实现:首先采用竞争聚集算法得到各数量型属性优化的模糊集个数,从而挖掘出优化的模糊关联规则。在得到用于构建预测系统规则库的模糊关联规则后,采用遗传算法约简冗余规则库,实现精确性和解释性的折衷,以提高预测精度。最后将此方法运用于 Abalone 样本数据集进行实验分析,证实此方法解决了模糊关联规则的冗余问题,有效提高了预测精度。

关键词:模糊关联规则;预测系统;竞争聚集算法;遗传算法
中图分类号:TP18 **文献标识码**:A **文章编号**:1001 - 4551(2010)06 - 0108 - 05

Optimization of fuzzy association rule based prediction system by genetic strategies

LI Yan, WANG Feng
(College of Computer, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Aiming at improving the accuracy of prediction system, an optimized method which based on fuzzy association rules was proposed for prediction system designing. The method was constructed by two phase, competitive agglomeration algorithm was employed to partition quantitative attributes from each data record into several optimized fuzzy sets, resulting in an initial prediction system. And a genetic algorithm was employed to optimize rule base, which can achieve a trade-off between accuracy and interpretability. This approach was applied to the Abalone data set, and demonstrated that this method solved the redundancy in fuzzy association rules and effectively improved the prediction accuracy.

Key words: fuzzy association rules; prediction system; competitive agglomeration algorithm; genetic algorithm

0 引 言

数据库内容丰富,蕴藏着大量可以做出智能商务决策和科学判断的信息,使用这些信息构建系统模型进行分类与预测一直是数据挖掘和机器学习核心研究内容之一。先前的研究主要侧重于分类问题,提出了很多分类系统的构造方法,如 C4.5、CBA、CFAR 等[1-3]。分类系统的输出值是离散的分类标签,如果输出属性的取值范围是连续论域,则变成了预测问题,它在电力负荷、水文监测、股票投资方面也有着广泛的应用价值。

通常预测问题都是通过建立合适的数学模型来解决,例如线性回归、多元回归、非线性回归和对数线性模型等,但是随着模糊集理论的出现,通过构建模糊系统的方法进行预测已经成为一种流行的途径,因为它具有开发周期短、不需要建立数学模型以及非线性等特点。构建模糊系统的关键问题是模糊规则的获取,随着数据库中关联规则挖掘技术的兴起,使用模糊关联规则挖掘算法挖掘有意义的模糊关联规则来构建基于模糊关联规则的预测系统(PFAR)的方法已经被提出。如文献[4]采用模糊 c 均值(FCM)算法划分数量型属性并挖掘模糊关联规

收稿日期:2009 - 10 - 22
基金项目:浙江省新苗人才计划项目(KYZ154308058 - 3)
作者简介:李 燕(1984 -),女,四川眉山人,主要从事智能控制和嵌入式技术等方面的研究. E-mail:chaoshangfeily@126.com

则来构建预测系统,具有较好的预测精度。但是,该方法的缺陷在于不知道将数量型属性划分成多少个模糊集比较合适,因而挖掘出的模糊关联规则并不能实现最优的预测精度。

本研究针对以上问题提出采用竞争聚集算法(CA)进行数据聚类[5],挖掘出优化的模糊关联规则,然后采用遗传算法来约简模糊规则集,实现解释性和精确性的折衷,解决模糊关联规则存在冗余问题影响预测精度的问题。

1　基于 CA 算法的模糊关联规则

1.1　应用 CA 算法对数据离散化

设 $T=\{t_1,\cdots,t_n\}$ 为一个关系数据库,t_j 表示 T 上的第 j 条记录,$I=\{i_1,\cdots,i_m\}$ 表示数量型属性集,$t_j[i_k]$ 表示第 j 个记录在属性 i_k 上的取值。模糊关联规则的挖掘首先需要采用 CA 算法将记录在各数量型属性上的值划分成若干优化的模糊集。

CA 算法是一种模糊聚类算法,它的基本思想是找到 c 个类的中心 $B=\{v_1,\cdots,v_c\}$ 以及 c 行 n 列的模糊划分矩阵 $(u_{ij})_{c\times n}$,使得计算得到的目标函数值最小:

$$J=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^m d^2(x_j,v_i)-\alpha\sum_{i=1}^{c}\Big[\sum_{j=1}^{n}u_{ij}\Big]^2 \quad (1)$$

式中　$u_{ij}\in[0,1]$,$\sum_{j=1}^{n}u_{ij}(1\leqslant i\leqslant c,1\leqslant j\leqslant n)$,$u_{ij}$—目标数据 x_j 隶属于第 i 个类的程度;m—大于 1 的模糊参数,用来控制聚类的模糊程度,通常取值为 2,当 $m\to1$ 时,聚类的划分趋于清晰,即 $u_{ij}\to1$ 或 $u_{ij}\to\infty$;当 $m\to\infty$ 时,聚类的划分趋于模糊,即 $u_{ij}\to1/m$;$d^2(x_j,v_i)$—目标数据 x_j 与类中心 v_i 之间的距离。

设目标数据集 $X=\{x_1,\cdots,x_n\}$,对 X 进行模糊聚类的 CA 算法迭代过程如下:

取定初始个数 $c=c_{\max}$,取定 ε,初始化迭代次数 $k=0$,初始化模糊 c 划分矩阵 $U^{(0)}$,计算聚类的基数:

$$N_i=\sum_{j=1}^{n}u_{ij}(1\leqslant i\leqslant c) \quad (2)$$

对 $k=0,1,\cdots$ 重复以下步骤:

(1) 计算 $d^2(x_j,v_i)(1\leqslant i\leqslant c,1\leqslant j\leqslant n)$;

(2) 修改 $\alpha(k)$ 的值;

$$\alpha(k)=\eta(k)\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^2d^2(x_j,v_i)}{\sum_{i=1}^{c}\big(\sum_{j=1}^{n}u_{ij}\big)^2} \quad (3)$$

$$\eta(k)=\eta_0\exp(-k/\tau) \quad (4)$$

其中,η_0 与 τ 根据具体情况取为某个固定常数,本研究取 $\eta_0=5$ 和 $\tau=10$。

(3) 修改划分矩阵 $U^{(k)}$;

$$u_{ij}=u_{ij}^{FCM}+u_{ij}^{Bias} \quad (5)$$

其中:

$$u_{ij}^{FCM}=\frac{[1/d^2(x_j,v_i)]}{\sum_{i=1}^{c}[1/d^2(x_j,v_i)]} \quad (6)$$

$$u_{ij}^{Bias}=\frac{\alpha(k)}{d^2(x_j,v_i)}(N_i-\bar{N}_j) \quad (7)$$

$\bar{N}_j$ 定义:

$$\bar{N}_j=\frac{\sum_{i=1}^{c}[1/d^2(x_j,v_i)]N_i}{\sum_{i=1}^{c}[1/d^2(x_j,v_i)]} \quad (8)$$

(4) 根据公式(1)重新计算聚类的基数 N_i,如果 $N_i<\varepsilon$,则丢弃此类。

(5) 修改聚类个数 c,利用矩阵 $U^{(k)}$ 中剩余的元素修改 v_i:

$$v_i=\sum_{j=1}^{n}(u_{ij})^2x_j\Big/\sum_{j=1}^{n}(u_{ij})^2(1\leqslant i\leqslant c) \quad (9)$$

(6) 置 $k=k+1$;

重复以上步骤直到中心的参数不再改变。

1.2　模糊集的表示

通过 CA 算法将 m 个数量型属性 $i_1,\cdots,i_m$ 上的取值根据数据实际分布情况划分成 $l_1,\cdots,l_m$ 个优化的模糊集。每个属性根据聚类中心的大小依次确定模糊等级,最大的中心对应最大的等级,以此类推。在模糊理论中,模糊集常采用三角模糊数、正态模糊数等模型来表示。为了简单起见,本研究采用三角模糊数表示模糊集,将模糊集表示成三角模糊数的方法如下。

假设 CA 算法对属性 i_k 上的数据集 X 聚类得到划分矩阵 U 和 c 个中心 v_k。记 $\mu_k(x_i)$ 是样本点 x_i 在第 k 个模糊集上的隶属度。将所有样本点根据最大隶属度原则归类,找出位于类中心 v_k 两侧的隶属度最小的样本点,设左侧隶属度最小的样本点为 x^l,隶属度为 $\mu_k(x^l)$,右侧隶属度最小的样本点为 x^r,隶属度为 $\mu_k(x^r)$,则第 k 个模糊集对应的三角模糊数 $f(x)$ 为:

$$f(x)=\begin{cases}\dfrac{x-a}{v_k-a} & a\leqslant x\leqslant v_k\\[2ex] \dfrac{b-x}{b-v_k} & v_k<x\leqslant b\\[2ex] 0 & a<x\text{ 或 }x>b\end{cases} \quad (10)$$

其中,$a=x^l-\dfrac{\mu_k(x^l)(v_k-x^l)}{1-\mu_k(x^l)}$;$b=x^r+\dfrac{\mu_k(x^r)(x^r-v_k)}{1-\mu_k(x^r)}$。

1.3 模糊关联规则的挖掘

通过 CA 算法得到原有数据库各数量型属性的模糊划分后,必须在此基础上构造一个新的数据库以进行模糊关联规则的挖掘,新数据库以数量型属性的不同模糊集等级作为数据库的属性,在此称为模糊属性。新数据库中记录在模糊属性上的取值方法为:不妨看模糊属性 $i_k(1)$,新数据库中第 j 个记录在模糊属性 $i_k(1)$ 上的取值为第 j 个记录在数量型属性 i_k 上的取值在模糊集 $i_k(1)$ 的隶属度,由此构造得到的新数据库含有 $l_1+\cdots+l_m$ 个模糊属性。不妨仍记所有模糊属性组成的集合为 I,第 j 个记录在模糊属性 y_k 上的取值为 $t_j(y_k)$,可知 $t_j(y_k)$ 属于区间[0,1]。

设 $X=\{y_1,y_2,\cdots,y_p\}$,$Y=\{y_{p+1},\cdots,y_{p+q}\}$,$X\cap Y=\Phi$,模糊关联规则以"$X=>Y$"的形式给出,其中规则前件 X 和后件 Y 中的模糊属性不应同时含有同一个 i_k 标记。挖掘模糊关联规则需要定义模糊支持率和模糊信任度,对于新构造的数据库,采用文献[4]类似的方法给出模糊支持率和模糊信任度的定义。

定义 1 模糊关联规则"$X=>Y$"的模糊支持率定义为 $FSup$:

$$FSup=\frac{\sum_{j=1}^{n}\prod_{m=1}^{p+q}t_j(y_m)}{n} \tag{11}$$

式中 n—T 的记录数;p—规则前件属性个数;q—规则后件属性个数。

定义 2 模糊关联规则"$X=>Y$"的模糊信任度定义为 $FConf$:

$$FConf=\frac{FSup}{\frac{1}{n}\sum_{j=1}^{n}\prod_{m=1}^{p}t_j(y_m)} \tag{12}$$

式中 n—T 的记录数;p—规则前件属性个数。

根据以上定义本研究可以采用类似于布尔型属性规则挖掘算法[6]挖掘不小于给定最小支持率和最小信任度的模糊关联规则作为预测系统规则库。

2 基于模糊关联规则的预测系统

一个基于模糊关联规则的预测系统主要由模糊关联预测知识库和模糊推理方法组成,如图 1 所示。

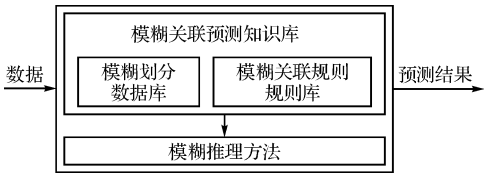

图 1 基于模糊关联规则的预测系统

模糊关联预测知识库中的规则库由上面提出的模糊关联规则的挖掘算法挖掘出的 M 条关联规则组成。规则形式为:

R_k:*If* i_1 *is* A_1^k *and* $\cdots$ *and* i_m *is* A_{m-1}^k, *then* y *is* A_m^k, $k=1,\cdots,M$.

其中,$A_1^k,\cdots,A_m^k$ 是数量型属性所取的三角模糊数表示的模糊集。而模糊划分数据库则由各数量型属性的模糊集定义,即前面提到的三角模糊数表示组成。模糊推理方法采用典型的 Mamdani 乘机推理机加重心解模糊化方法[7]。

3 模糊预测系统的遗传优化

模糊预测系统的精度和精简程度是其两个重要指标,在基于模糊关联规则的预测系统中,精简程度用构建预测系统的模糊关联规则的数目来衡量,数目越少,系统就越精简。然而采用模糊关联规则挖掘算法时由于最小支持率和最小信任度的设定问题难免会挖掘出许多冗余的规则,这样不仅系统解释性不高更会影响预测精度。为了解决此问题,下面采用遗传算法进行优化,使其达到精确性和解释性的折衷。优化过程通过以下二进制编码遗传算法实现。

(1) 编码。

对于规则库的编码本研究采用匹兹堡方法[8]。考虑具有 m 条规则的初始规则,规则记为 $R_i(i=1,\cdots,m)$,那么可以用一个 m 位的二进制位串(染色体)$S=(s_1,\cdots,s_m)$ 表示形成最终规则库 B^F 的候选规则集的一个子集,它满足:如果 $s_i(i=1,\cdots,m)$ 为 1,则 $R_i\in B^F$;否则 $R_i\notin B^F$。

(2) 初始种群的产生。

初始种群的第一个个体通过引入一个表示完整的初始规则集的染色体来生成,即每个基因 s_i 都等于 1。其余的个体随机生成,即每个基因随机地取 0 或 1。

(3) 适应度函数的设定。

设初始规则集的一个子集 S,其所含规则数为 $N(S)$,预测绝对误差和为 $E(S)$。那么,简化的目标可以考虑为 $N(S)$ 和 $E(S)$,这是一个两目标的组合优化问题,引入权值 $0<\omega<1$,则定义规则集 S 的适应度 $f(S)$ 为:

$$f(S)=\begin{cases}\omega\dfrac{E(S)}{E_0}+(1-\omega)\dfrac{N(S)}{N_0} & E_0\neq 0\\ \omega E(S)+(1-\omega)\dfrac{N(S)}{N_0} & E_0=0\end{cases} \tag{13}$$

式中 E_0—使用初始规则库预测的绝对误差;N_0—初始规则库包含的规则数目。

对于当代种群中的个体都采用上述适应度函数进行评估。

(4) 遗传操作。

遗传操作包括选择、交叉、变异,它能控制着种群向最优解收敛。本研究的选择操作采用轮盘赌方法,使得适应度越大的个体被保留到下一代的几率越大。交叉操作使用标准的二进制两点交叉操作生成下一代个体群。两点交叉操作类似于单点交叉,只是随机地设置两个交叉点。当两个父代个体进行交叉时,在两个交叉点之间的码串相互交换,其他基因座的码串不变,从而分别生成两个新的后代个体。对由变异操作生成的各个体,按变异概率进行基本变异操作。变异的基因由 0 变为 1 或由 1 变为 0。另外需要说明的是对交叉和变异操作都按固定概率进行,即假设种群大小为 N,编码长度为 m,交叉率 P_c,变异率 P_m,则每一代都随机选择 $N \times P_c$ 个个体进行交叉操作,随机选择 $N \times m \times P_m$ 个基因进行变异操作。

4　实例分析

为了检验此方法的有效性,本研究选用 UCI Machine Learning Repository 上的预测数据集 Abalone 进行实验分析。Abalone 数据库含有 8 个数量型属性分别即为 $i_1,\cdots,i_8$,实验根据数据库中前 7 个属性 $i_1,\cdots,i_7$ 的取值来预测数量型属性 i_8(年龄)的值。其中的一份测试样本如表 1 所示。

表 1　Abalone 测试样本

属性	样本取值	属性	样本取值
长度	0.455	脱壳重	0.224 5
直径	0.365	内脏重	0.101
高度	0.095	壳重	0.15
全重	0.514	年龄	15

实验中笔者将 4 177 条记录分成两组,分别为训练数据集和测试数据集。训练数据集用来进行模糊关联规则的挖掘以及在此基础上的遗传算法适应度值的评估,而测试数据集合则用来测试预测精度。规定 70% 的记录用作训练数据,30% 的记录用作测试数据。本研究采用 CA 算法对 8 个数量型属性进行模糊划分后得到的结果为:3,3,2,4,3,3,3,3。设定最小支持率为 0.2%,最小信任度为 50%,采用模糊关联规则的挖掘算法挖掘得到了 207 条规则。接着采用遗传算法约简冗余规则库,初始种群个数设为 20,交叉率和变异率分别设定为 0.6 和 0.003,权值参数 ω 设定为 0.85。随着遗传算法的推进,在测试样本的平均线性误差以及优化得到规则数目的变化趋势如图 2、图 3 所示。

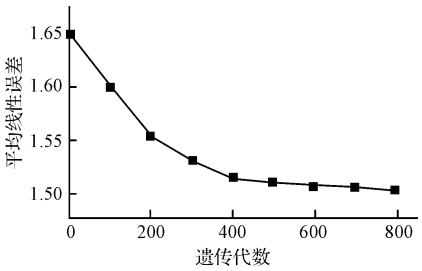

图 2　平均线性误差随遗传代数变化趋势

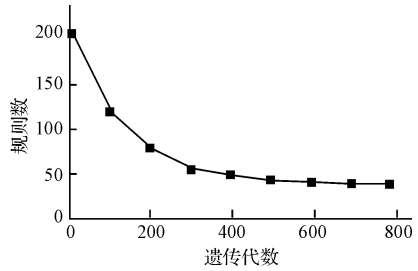

图 3　规则数随遗传代数变化趋势

实验结果显示:随着迭代次数的增加,测试样本平均线性误差逐渐变小,规则数目逐渐减少,并趋于稳定。另外,基于同样的数据集,采用 FCM 算法划分数量型属性构建的模糊预测系统以及未采用遗传算法约简规则库的模糊预测系统对测试数据的实验室结果如表 2 所示。比较结果表明,采用本研究的方法可以得到相比其他方法更好的预测精度,同时系统也达到了很好的精简度。

表 2　几种方法的比较

方法	平均线性误差	规则数
PFAR(FCM)	1.752 5	220
PFAR(CA)	1.647 5	207
PFAR(CA&GA)	1.504 7	37

5　结束语

本研究提出了一种优化的模糊预测系统设计方法,首先通过 CA 算法划分得到各属性优化的模糊集从而挖掘优化的模糊关联规则,在此基础上为了实现精度和精简程度的折衷,采用遗传算法优化规则库。实验结果表明采用本方法可以进一步提高预测系统的精度,并且系统更加精简。

(下转第 123 页)

权重系数按与 x 点的距离赋值。

经二值化和八邻域搜索得到轮廓后,对轮廓即可作傅里叶变换。为使傅里叶描述子的误差尽量减少,作傅里叶变换的轮廓图像应该大小一致。同时计算时需要兼顾计算的复杂度和精确度,经过经验对比可知,对一个轮廓采用 12 个描述子是较为合适的。

MEB-SVM 算法中的 ε 是决定算法精确度与复杂度的关键因素。实验中分别取不同的 ε 进行训练,得到的结果如表 1 所示。

表 1　实验结果

ε	训练时间/s	识别准确度/%
0.000 5	3.07	80.2
0.000 8	4.11	86.5
0.001	5.29	91.4
0.001 4	7.18	92.2
0.002	10.26	93.6
0.002 5	15.37	94.5

实验结果中,ε 越小,算法收敛时间越短,但由于包含在包围球中的向量太少,对算法的准确度会有一定的影响,当 ε 大于 0.001 以上时,算法训练时间增加明显,同时算法准确度也有所增加。

5　结束语

本研究介绍了一种基于 MEB-SVM 的手势识别算法,MEB-SVM 算法是在具有强分类能力的 SVM 的基础上提出来的,且此方法的计算复杂度与精确度又超过了传统的 SVM 方法,具有非常广泛的应用前景。

参考文献(References):

[1] ATID S, WU H, ALISTAIR S. Hand gesture recognition for human computer interaction[N]. ERCIM NEWS, European Research Consortium for Informatics and MAthematics,2001(46).

[2] WU J Q, GAO W, PANG B, et al. A fast sign word recognition technique for Chinese sign language[J]. **High Technology Letters**,2001,11(6):23 - 27.

[3] CERVANTES J, LI Xiao-ou, YU Wen, et al. Support vector machine classification for large data sets via minimum enclosing ball clustering[J]. **Neurocomputing**,2008,71(4 - 6):611 - 619.

[4] CHENG Yi-zong. Mean shift, mode seeking and clustering [J]. **IEEE Trans on Pattern Analysis and Machine Intelligence**,1995,17(8):790 - 799.

[5] CORTES C, VAPNIK V N. Support-vector networks[J]. **Machine Learning**,1995,20:273 - 297.

[6] BOSER B, GUYON I, VAPNIK V N. A training algorithm or optimal margin classifiers[C]//Proc. Fifth Annual Workshop on Computational Learning Theory. New York: [s. n.],1992:144 - 152.

[7] 段一洪,陈一民,林　锋. 基于 LSSVM 的静态手势识别[J]. 计算机工程与设计,2004,12(25):2352 - 2368.

[8] TSANG I W, KWOK J T, CHEUNG P M. Core vector machines: fast SVM training on very large data sets[J]. Journal of Machine Learning Research,2005,6:363 - 392.

[编辑:李　辉]

(上接第 111 页)

参考文献(References):

[1] QUINLAN J R. C4.5: programs for machine learning[M]. San Mateo: Morgan Kaufmann,1993.

[2] LIU Bing, HSU W, MA Yi-ming. Integrating Classification and Association Rule Mining[C]//Proc. of the International Conf. on Knowledge Discovery and Data Mining. New York: [s. n.],1998: [s. n.].

[3] LU J, XU B, KANG D. Classification methods of association rules with linguistic terms[J]. **Journal of Southeast University**,2004,20(1):21 - 25.

[4] LU J, XU B, JIANG J. A Prediction Method of Fuzzy Association Rules[C]//Proc. of the International Conf. on Information Reuse and Integration. Nevada: [s. n.],2003: [s. n.].

[5] FRIGUI H, KRISHNAPURAM R. Clustering by competitive agglomeration[J]. **Pattern Recognition**, 1998, 30(7): 1109 - 1119.

[6] AGRAWAL R, SKIKANT R. Fast Algorithms for Mining Association Rules[C]//Proc. of 20th International Conf. Very on Very Large Databases. Morgan Kaufmann: [s. n.], 1994: [s. n.].

[7] 王立新,王迎军. 模糊系统与模糊控制教程[M]. 北京:清华大学出版社,2003.

[8] SMITH S F. A learning system based on genetic algorithms [D]. Pittsburgh: University of Pittsburgh,1980.

[编辑:李　辉]