*Hans Journal of Data Mining* 数据挖掘, 2012, 2, 33-37
http://dx.doi.org/10.12677/hjdm.2012.24007 Published Online October 2012 (http://www.hanspub.org/journal/hjdm.html)

# Research on Associative Classification Method Based on Threshold Adaptive Technology*

**Wei Tang, Jingxue Liu**

PLA Academy of National Defence Information, Wuhan
Email: ljx_62@sohu.com

Received: Jul. 26th, 2012; revised: Aug. 10th, 2012; accepted: Aug. 24th, 2012

**Abstract:** Based on the mechanism of evaluation feedback and control, this paper researches for adaptive adjustment of setting threshold of associative classification method. Firstly, the mathematical model of evaluating the set that consist of classification rules is established. Then, based on this model, a method that uses the interval iterative approximation for getting the approximate optimal solution of associative classification minimum support in order to achieve the adaptive threshold settings is put forward. Finally, the simulation experiment shows that the method has better adaptability and effectiveness comparing with some general ones.

**Keywords:** Self-Adaptive; Associative Classification; Rule Extraction; Class Being Distinguished

# 阈值自适应关联分类方法研究*

**汤　伟，刘敬学**

国防信息学院，武汉
Email: ljx_62@sohu.com

收稿日期：2012 年 7 月 26 日；修回日期：2012 年 8 月 10 日；录用日期：2012 年 8 月 24 日

**摘　要：**本文采用评估反馈控制机制，对关联分类方法的阈值设定自适应调节进行了研究。首先建立了对分类规则集进行判优评估的数学模型；然后基于此模型提出了运用区间迭代逼近求解关联分类最小支持度，并以此进行阈值自适应调节的方法；最后的仿真实验表明，该方法比一般的关联分类方法具有更好的自适应性和有效性。

**关键词：**自适应；关联分类；规则提取；类型判别

## 1. 引言

基于关联规则挖掘(Frequent Pattern Mining)的分类方法称为关联分类(Associative Classification)[1]，用于基于带类标记的训练数据集进行分析，而不是传统的事务数据集，规则应用的重点也由原来的提取知识变为对新数据的分类和趋势预测。而用于关联分类的训练数据集具有两类不同属性[2]，一类是组成分类规则前件(条件)的条件属性，记为 $A_1,A_2,...,A_n$，另一类是组成后件(结论)的决策属性，记为 $C_1,C_2,...,C_m$。决策属性的值预先标记多属性目标的类别，类别的标记可由多个决策属性联合标记来完成。

关联分类主要应用于类型判别系统 $DS = (A \cup C,R,M)$，其中 $R$ 为分类规则集，$M$ 为分类器的构建方法。实际应用中，系统利用分类器对测试数据进行类型判别，判别的准确率依赖于 $R$ 的可靠度和 $M$ 的优劣。传统的关联分类采用经验知识或人工尝试的办法来寻找合适的支持度和置信度阈值，明显存在效率不高、智能性不强的问题。为了提高分类规则的可靠性，

*资助信息：本研究课题受到国家自然科学基金(70903026)资助。

 

本文基于分类规则集的综合评估指标值智能调节自身阈值设定的反馈自适应控制机制，探讨对关联分类方法的阈值设定进行自适应调节的改进和具体实现方法，以提高分类决策的智能性与精准度，降低其对噪声数据和孤立点的敏感性。

在本文第 2 部分，对建立分类规则集判优评估的数学模型进行了阐述；在第 3 部分，先分析了阈值自适应关联分类方法实现的基本思路，然后基于判优评估模型提出了区间迭代逼近求解关联分类最小支持度阈值的方法，从而实现了阈值的自适应调节；在第 4 部分，通过仿真实验，验证了阈值自适应关联分类方法的智能性和有效性。

## 2. 分类规则集判优评估模型的建立

关联分类挖掘的数据集属性值如果是连续的，须先对其进行离散化处理，设 $v$ 表示某个属性值处理后的单个取值或离散区间的标号，先对数据集条件属性值标签化处理，转换为关联挖掘的一个项 $p$，$p=(A_i,v)$，再将决策属性 $C_1,C_2,...,C_m$ 联合成分类属性 Aclass，记 $C$ 为其标识的一个类，则分类规则 $r$ 表示为如下形式的蕴含式[1]：

$$p_1 \wedge p_2 \wedge ... p_l \Rightarrow A_{\text{class}} = C,\ (1 \le l \le n) \qquad (1)$$

关联分类的主要任务是运用关联挖掘算法挖掘形如式(1)的分类规则，其实质是一类特殊的频繁项集(Frequent Item Set)，可采用剪枝的方法和预设阈值产生强规则，最后对规则进行粗处理得出分类规则集。

将训练数据集分割为基本训练集和测试训练集两个部分，基本训练集用来挖掘分类规则，测试训练集用来测试评估分类规则的优劣，以进行阈值的反馈控制调节。

可以对训练数据集采用一次划分或 $k$-折划分的方法。在一次划分法中，训练数据集按比例划分成两部分，测试训练集的综合测试评估指标值 $Z$ 作为目标函数；在 $k$-折划分法中，训练数据集被划分成大小大致相等的 $k$ 个互不相交的子集 $S_1,S_2,...,S_k,(k\geq 2)$，训练进行 $k$ 次，在第 $i$ 次训练中，$S_i$ 用作测试训练集，其余子集的集合用作基本训练集，设最小支持度阈值为 $ms$，最小置信度阈值为 $mc$，第 $i$ 次测试训练的评估指标值为 $Z_i$，则综合判优评估指标 $Z(ms,mc)$ 用算术平均法求得，即：

$$Z(ms,mc) = \sum_{i=1}^{k} Z_i \Big/ k,\ (k \ge 1) \qquad (2)$$

其中，$k$ 表示划分的折数，一般建议使用 10-折划分，因为它具有相对低的偏置和方差，特别地，当 $k = 1$ 时，表示使用的是一次划分法。

考虑到数据的多样化和对象并不都是唯一可分类的，采用对规则评判而不是对数据测试的评判更具有合理性。把基于设定的最小阈值训练产生的一组强关联分类规则记为 $R=\{r_1,r_2,...,r_m\}$，$r_i$ 的支持度和置信度分别记为 $s_i$ 和 $c_i$。

对规则 $r_i$ 进行测试，记该规则的判别前件为 $A(r_i)$，判别类型为 $C(r_i)$，满足规则前件和结论的数据记录个数分别为 $N(A(r_i))$ 和 $N(C(r_i))$，测试数据集总的记录数为 $N$。测试准确率表示为 $n_i=L_i/P_i$，其中 $n_i$ 为准确率，$L_i$ 为判别正确的记录数，$P_i$ 为满足规则前件的数据记录数；测试误检率表示为 $f_i = M_i/N_i$，其中 $f_i$ 为误检率，$M_i$ 为判断错误的记录数，$N_i$ 为非 $C(r_i)$ 类型的数据记录数。以这四项指标作为规则集的评判属性，得决策矩阵：

$$\boldsymbol{D} = \begin{pmatrix} s_1 & c_1 & n_1 & f_1 \\ s_2 & c_2 & n_2 & f_2 \\ \vdots & \vdots & \vdots & \vdots \\ s_m & c_m & n_m & f_m \end{pmatrix} \qquad (3)$$

对规则的测试而言，$s_i$、$c_i$、$n_i$ 要求越高越好，而 $f_i$ 要求越低越好，运用乘除法综合式(3)中的效益型和成本型属性，得评估规则测试指标 $T_i$：

$$\begin{aligned} T_i &= c_i \cdot n_i \cdot (1-f_i) = c_i \cdot \frac{L_i}{P_i} \cdot \left(1-\frac{M_i}{N_i}\right) \\ &= \frac{c_i \cdot L_i \cdot \left(N\left(\overline{C}(r_i)\right) - M_i\right)}{N\left(A(r_i)\right) \cdot N\left(\overline{C}(r_i)\right)} \end{aligned} \qquad (4)$$

其中 $N\left(\overline{C}(r_i)\right) = N - N\left(C(r_i)\right)$。

置信度高的规则具有更高的可靠性，在规则集综合评估中的权重也应该越大，故可以运用简单加权平均方法进行评估，它是一种非等权平均方法。按置信度大小对规则进行排序得到新的决策矩阵 $\boldsymbol{D}'$，使得 $s_1 \leqslant s_2 \leqslant \cdots \leqslant s_m$，令新规则序列中 $r_j$ 的权系数 $w_j$ 为：

$$w_j = \frac{j}{\sum_{j=1}^{m} j} = \frac{2j}{m(m+1)} \qquad (5)$$

记规则集的所有判别条件的集合为 $\bigcup_{r\in R} A(r)$，其覆盖度定义为 $N\left(\bigcup_{r\in R} A(r)\right)\Big/N$；当规则集为空时，其覆盖度为零。综合加权测试指标与覆盖度指标，可得第 $i$ 次测试训练的评估指标 $Z_i$：

$$Z_i = \sum_{j=1}^{m} w_j \cdot T_j \cdot N\left(\bigcup_{r\in R} A(r)\right)\Big/N \tag{6}$$

将式(2)与式(6)合并，得到规则集的综合判优评估模型为：

$$\max Z(ms,mc) = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{w_j \cdot T_j \cdot N\left(\bigcup_{r\in R} A(r)\right)}{k\cdot N} \tag{7}$$

$ms$ 用于筛选出关联度强的规则。$ms$ 过高会导致得出的分类规则集不能完整描述各类别的推理特征；$ms$ 过低会导致得到一些可靠性较低的规则，误导判别结果，也会使得出的规则过于繁杂。而 $mc$ 用于筛选出可信度高的规则，越高则基于准确率和误检率的综合判优评估指标值会越高，但片面地要求较高的指标值会导致得出的规则集数目过少，有用的规则被剪枝。

因此，在实际应用中，可先确定 $mc$ 的初值，以求出最优的支持度阈值解，在此基础上给出 $mc$ 的推荐取值，最后由决策者根据实际要求确定 $mc$ 的取值。

此外，分类规则集应能涵盖至少一种类别的特征，$ms$ 需小于最大数目类别所占总数的比例，设为 $p\%(0 < p < 100)$，训练样本总数设为 $N$，将式(4)，式(5)代入模型(7)，修正并化简为：

$$\max Z = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{2j\cdot c_j \cdot L_j \cdot \left(N\left(\overline{C}\left(r_j\right)\right) - M_j\right)\cdot N\left(\bigcup_{r\in R} A(r)\right)}{m\cdot(m+1)\cdot k\cdot N\left(A\left(r_j\right)\right)\cdot N\left(\overline{C}\left(r_j\right)\right)\cdot N}$$

$$s.t.\begin{cases} 0 < ms \le 0.01p \\ 1 \le k << N \end{cases} \tag{8}$$

这样，原问题转换为求解一个最优的最小支持度阈值解 $ms^*$，使得 $Z(ms)$达到最大。

## 3. 阈值自适应关联分类方法的实现

### 3.1. 基本思路

阈值自适应关联分类方法的实现主要是根据所得分类规则集的评估指标值智能调节阈值，包括对训练数据属性值进行预处理、基于评估反馈控制机制求出的最小支持度阈值产生分类规则、基于已产生的分类规则构建分类器、利用分类器判别目标类型等步骤[6]。具体实现流程如图 1 所示。

### 3.2. 评估反馈控制的区间迭代逼近算法

为了能求解出式(8)的近似最优解，本文采用一种基于评估反馈控制的区间迭代逼近求解方法。该算法设置区间迭代指数为一个大于 1 的正整数 $e$，$ms$ 的取值区间为$(a,b]$；具体步骤如下：

Step1 输入训练数据集 $S$，最小置信度阈值 $mc$，区间迭代指数 $e$；将训练数据集按某种划分方法进行划分。

Step2 将 $ms$ 的取值区间划分为连续等长的 $e$ 个子区间$(a_{i-1},a_i]$，$(i=1,2,...,e)$，取各子区间的中点$(a_{i-1}+a_i)/2$，设为 $b_i$。

Step3 分别取 $ms = b_i$ 对数据集进行基本训练和测试训练，得综合判优评估指标值序列$(Z(b_1),Z(b_2),\cdots,Z(b_e))$。

Step4 找出最大的评估指标值 $\max Z$，对应的 $ms$ 值为 $b_j$。

Step5 若 $j = 1$ 或 $j = e$，将 $ms$ 的取值区间分别缩小为$[a,b_2]$和$[b_{e-1},b]$；否则将取值区间缩小为$[b_{j-1},b_{j+1}]$。

Step6 转 Step2，直到准则函数

$$E = \sum_{i=1}^{e}\left|Z\left(b_i\right) - \sum_{j=1}^{e} Z\left(b_j\right)\Big/e\right|^2$$

收敛，得近似最优解 $ms^*=(a+b)/2$。

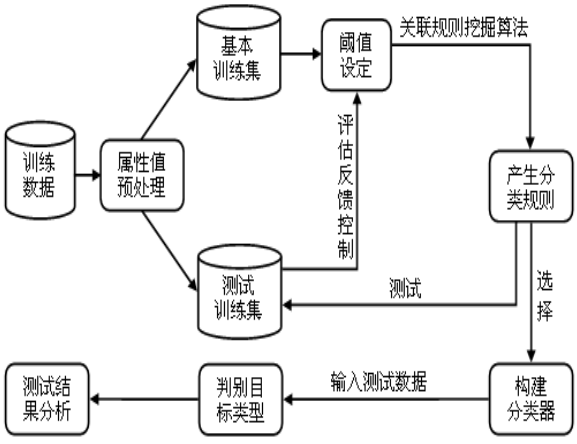


**Figure 1. flow chart of associative classification based on self-adaptive threshold**
**图 1. 阈值自适应关联分类方法实现流程图**

Step7 输出近似最优解 $ms^*$和分类规则集。

从第 2 部分建立的模型来看，支持度阈值越偏离某一理想小区间，评估值越低，且呈现先单调递增后递减的特点；因此，上述算法是一个能收敛到最优解的逼近算法，其收敛速度与 $e$ 等迭代指数密切相关。

在求解出 $ms^*$的基础上，决策者再根据具体要求调整 $mc$，以使得出的分类规则集达到最可靠状态。但最终判别目标类别还依赖于采用具体的方法构建分类器。

## 4. 仿真实验

对阈值自适应关联分类方法的仿真实验，基于.NET平台架构，采用算法Apriori[4]作为关联规则的挖掘算法，分别选用文献[5]中的三个数据集进行实验。数据集的描述如表 1 所示。

对这三个数据集，采用一次划分法进行分割，区间迭代指数取值 $e$ = 16。按照阈值自适应区间迭代逼近求解方法，求解过程分别迭代了{7,7,6}次，求得最小支持度近似最优解 $ms^* \approx$ {11.25%,7.58%,0.67%}。

在初始取值区间{(0,0.3375)，(0,0.3864)，(0,0.3333)}之间选取{33,38,33}个阈值采样点，这些采样点与综合判优评估指标 $Z$ 值的关系如图 2 所示。

对关联分析产生的分类规则采用启发式方法构建分类器，并测试其对目标类型判别的正确率，测试结果与文献[6]中提出的基于排序的关联分类算法作了比较，具体如表 2 所示。

通过以上两种方法分别对三类数据集的测试、分析和比较实验表明，由本文提出的逼近算法求得的近似最优解构建的分类器有较高的测试正确率，能分析出最为有效的分类规则，且计算程序的运行过程和输出结果进一步验证了区间迭代逼近求解方法的可行性，进而验证了阈值自适应关联分类方法的智能性和可靠性，仿真结果与理论分析相一致。

**Table 1. Description of the data set**
**表 1. 数据集描述**

| 数据集 | 记录数 | 条件属性个数 | 决策属性个数 | 类别数目 | 属性值中是否有缺省值 |
|---|---|---|---|---|---|
| Acute Inflammations | 120 | 6 | 2 | 4 | 否 |
| Hayes-Roth | 160 | 3 | 1 | 3 | 否 |
| Iris | 150 | 4 | 1 | 3 | 否 |

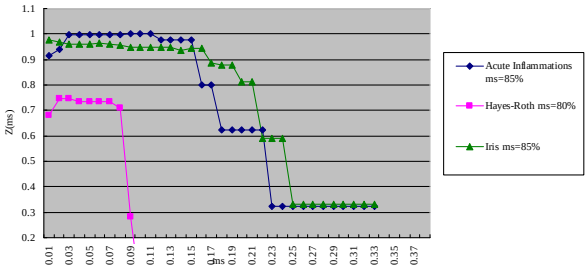


**Figure 2. Relation chart of the threshold's sample points corresponding to the values of synthetically evaluating**
**图 2. 阈值采样点与综合判优评估指标值对应关系图**

**Table 2. Test result's contrasts of the two kinds of associative classification method**
**表 2. 两种关联分类方法测试结果对比**

| 分类方法 | 测试数据集 | 支持度阈值 | 分类规则数目 | 测试数据数目 | 测试正确率 |
|---|---|---|---|---|---|
| 基于排序的关联分类方法 | Acute Inflammations | 20% | 28 | 120 | 84.17% |
| | Hayes-Roth | 5% | 15 | 80 | 91.25% |
| | Iris | 3% | 75 | 150 | 94.67% |
| 阈值自适应关联分类方法 | Acute Inflammations | 11.25% | 136 | 120 | 100% |
| | Hayes-Roth | 7.58% | 9 | 80 | 91.25% |
| | Iris | 0.67% | 140 | 150 | 96% |

## 5. 结束语

本文针对一般关联分类方法存在的依赖先验知识和智能性不强、准确度不够高等缺陷，提出并实现了基于评估反馈控制机制的阈值自适应关联分类方法，并选用三个数据集进行了仿真实验，仿真结果与理论分析相一致，验证了该方法比一般关联分类方法有更好的自适应性和有效性。该方法在目标类型识别中具有较高的应用价值，可辅助提高决策的准确率和对先验知识的利用率。下一步值得深入研究的问题有：一是随着属性数和数据量的增大，如何提高算法的效率；二是在判优评估模型中，综合考虑速度、鲁棒性、可规模性和可解释性等指标的度量，进一步提升其应用价值。

## 6. 致谢

本文得到了国家自然科学基金(70903026)的资助和国防信息学院的支持。

## 参考文献 (References)

[1] J. W. Han, M. Kamber. Data mining concepts and techniques.

2nd Edition, 北京: 机械工业出版社, 2006: 344-345.
[2] 丛蓉. 作战指挥决策支持系统目标融合识别研究[D]. 大连理工大学, 2010: 93-98.
[3] 晁玉宁, 许孝元. 基于关联规则的分类模型系统[J]. 计算机工程与应用, 2009, 7: 80-83.
[4] H. Wu, Z. G. Lu, L. Pan, et al. An improved apriori-based algorithm for association rules mining. Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009: 51-55.
[5] D. Aha, Fellow Graduate Students. UCI machine learning repository. Irvine: University of California. http://archive.ics.uci.edu/ml/datasets.html
[6] 朱晓燕, 宋擒豹. 基于排序的关联分类算法[J]. 计算机科学, 2009, 36(7): 204-207.

 