第45卷 第9期 西 安 交 通 大 学 学 报 Vol. 45 No. 9
2011年9月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Sep. 2011

# 具有全局聚类的多属性离散化算法

刘弹，杨景明，罗爱玲
（西安交通大学机械制造系统工程国家重点实验室，710049，西安）

**摘要**：为了减少连续属性离散化后有用信息的丢失和信息系统总的断点数量，提出了一种具有全局聚类效果的多属性离散化算法.算法根据各属性预插入断点对信息系统近似分类质量的影响，来确定要插入断点的属性，从全局属性范围选择最佳断点.根据Ameva统计量来判断属性中最佳断点的位置，并以保证决策表的近似分类质量作为算法的终止条件.实验采用多组机器学习数据对算法的性能进行了检验，并与几种经典算法做了对比.实验结果表明，用新的离散化算法获得的结果所建的C45决策树分类模型，具有较好的分类精度和较少的节点数量.

**关键词**：统计量；连续属性；离散化
**中图分类号**：TP18 **文献标志码**：A **文章编号**：0253-987X(2011)09-0001-05

## Synchronized Continuous Attributes Discretization Based on Ameva

LIU Dan, YANG Jingming, LUO Ailing
(State Key Laboratory for Manufacturing Systems Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

**Abstract**: To avoid information loss and cut points decrease after discretization of continuous attributes, a synchronized continuous attribute discretization algorithm with good global clustering effect for selecting cut points from all conditions attributes is presented. This algorithm decides which continuous attribute should be inserted according to the cut point from all attributes based on the influence of the inserted cut point. The influence is evaluated by information system approximation classification quality. Then cut point is selected from the candidate points in the attribute according to Ameva statistics, and the level of indiscernibility relation is chosen as the stopping condition of the algorithm. By UCI machine learning data sets a comparison with several classic discretization algorithms shows that the C45 classification model based on the proposed algorithm is of good classification accuracy and needs less nodes.

**Keywords**: statistics; continuous attributes; discretization

机器学习和数据挖掘的许多算法都要求属性是离散值，而实际获取的数据却是连续属性，因此需要对连续属性进行离散化处理[1].作为数据预处理的重要步骤，连续属性离散化的效果直接影响了学习的过程和学习的最终结果[2].离散化本质是为了利用适合的断点数量和断点位置对条件属性构成的空间进行划分.

为解决连续属性离散化的问题，研究人工智能的学者提出了很多离散化的算法，而离散化算法按照不同的分类标准可以分为有监督与无监督、拆分式与合并式、直接式与增量式、局部离散化与全局离散化等类型.其中，基于皮尔逊统计量的方法是最有影响力的方法之一.1992年Kerber[3]应用统计学的思想，提出了ChiMerge启发式的连续属性离散化算法，该算法应用统计学中的皮尔逊统计量来度量属性上的划分样本分类之间的关系，并用以判别断

收稿日期：2011-03-08. 作者简介：刘弹(1978－)，男，讲师. 基金项目：国家自然科学基金资助项目(51105296)；机械制造系统工程国家重点实验室开放课题资助项目；中央高校基本科研业务费专项资金资助项目.
网络出版时间：2011-07-29 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20110729.1609.001.html

点相邻区间是否可以合并. 在此基础之上,Liu 等人[4-6]又提出了 Chi2、Modified Chi2、Extended Chi2 等改进算法. 1994 年 Marce[7]通过皮尔逊统计量,以最小化待离散的条件属性和类属性之间的置信水平为依据,提出了 Khiops 算法,解决了 ChiMerge 算法中统计检验不可靠的问题. 2009 年 Gonzalez 等人[8]以皮尔逊统计量为基础,引入区间和类别参数对独立性检验的参数做出适当的调整,提出了 Ameva 算法,该算法不需要提供任何的阈值和参数,并且在减少离散化后的区间数量上获得了良好的效果.

以上算法是从数据在单个属性上的分布来考虑的,而不是整个属性空间上的分布. 因此,算法寻求的是数据集的局部最优解,而不是从整个数据集的角度来求解,离散化的结果必然存在不合理的划分断点. 本文以 Ameva 算法的研究为基础,考虑到单属性离散化的不足,结合粗糙集理论所挖掘的信息,将算法从单属性范围拓展到多属性,提出了一种多属性离散化算法. 该算法对多个条件属性进行并行离散处理,旨在从全局范围内搜索最佳离散断点,离散过程中综合考虑了各属性之间数据的分布,从而使数据获得了良好的全局聚类效果,并同时保证了决策表的近似分类质量.

## 1 粗糙集理论的基本概念[9]

在粗糙集理论中,决策表是一个由四元组构成的信息表知识表达系统$(U,A,V,f)$,其中论域$U$是对象的集合,即$A=C\cup D$是属性的集合,子集$C$和$D$分别是条件属性集合和决策属性集合,$V$是由属性的取值范围构成的集合,$f:U\times A\to V$是一个信息函数,它指定$U$中的每一个对象在各个属性上的取值,$D\neq\varnothing$.

**定义 1** 集合的上下近似集. 令属性集$R\in A$,它的不可分辨关系为

$$\left.\begin{aligned} I_R=(x,y)\in U\times U\mid\ &\forall a\in R\\ &f(x,a)=f(y,a)\end{aligned}\right\}\tag{1}$$

以$U/I_R$表示$I_R$的所有等价类构成的集合. 对于$X\subseteq U$,$R\subseteq A$,$X$关于$R$的下近似集合为

$$R\_(X)=\bigcup\{Y\in U/R\mid Y\subseteq X\}\tag{2}$$

$X$关于$R$的上近似集为

$$R^{-}(X)=\bigcup\{Y\in U/R\mid Y\cap X\neq\varnothing\}\tag{3}$$

**定义 2** 近似分类质量. 设$\pi(U)=\{x_1,x_2,\cdots,x_n\}$是论域$U$的一个划分,且这个划分独立于属性集$R$. 其中,子集$x_i(i=1,2,\cdots,n)$是划分$\pi(U)$的等价类. 划分$\pi(U)$的$R$的近似分类质量为

$$\gamma_R(\pi(U))=\frac{\sum_{i=1}^{n}\mid R\_(x_i)\mid}{\mid U\mid}=\frac{\mid R\_(\pi(U))\mid}{\mid U\mid}\tag{4}$$

## 2 基于 Ameva 的多属性离散化算法

$(U,A,V,f)$中的$U=\{x_1,x_2,\cdots,x_n\}$,对每一个连续条件属性$a\in C$,设其值域为$V_a=[l_a,r_a]$. 若$a$上有一组点$l_a<c_1^a<c_2^a<\cdots<c_{m_a}^a<r_a$,则这一组点按照区间划分将$a$的取值划分成$m_a+1$个等价类,因此将每一个$c_i^a$定义为一个断点.

$a$在论域中的有限个属性值经过排序后为

$$l_a=v_1^a<v_2^a<\cdots<v_{n+1}^a=r_a$$

则候选断点可取为

$$c_i^a=(v_i^a+v_{i+1}^a)/2,\ i=1,2,\cdots,n\tag{5}$$

用$c_i^a$来描述属性$a$的第$i$个断点$(1\leqslant i\leqslant n)$,$n$为属性$a$的断点总数,集合$X(X\subseteq U)$是由断点$c_i^a$分开的实例集合.

单属性离散化算法趋向于选择具有最大决策属性值辨别能力的点作为划分的断点[10],如表 1 所示,$a$、$b$表示条件属性,$d$表示决策属性. 在使用单属性离散化算法对$a$做离散处理时,根据数据分布,为取得较好的聚类效果,需要插入断点为 2.85 的数,才能以最大概率将具有不同决策属性值的实例分开. 同样对$b$做离散处理时,则需插入断点为 2.1 的数.

表 1 原始决策表

| $U$ | $a$ | $b$ | $d$ |
|---|---|---|---|
| 1 | 1.6 | 1.0 | 1 |
| 2 | 2.4 | 1.7 | 1 |
| 3 | 3.3 | 2.5 | 0 |
| 4 | 3.6 | 5.0 | 0 |
| 5 | 3.8 | 1.5 | 1 |
| 6 | 3.8 | 4.0 | 0 |
| 7 | 4.4 | 1.5 | 0 |

从表 2 中可以发现,单属性离散化后的决策表存在不一致的实例,图 1 为对单属性离散化图形的描述. 对于整个决策系统来说,经过单属性离散化后,整个决策表被划分为 4 块. 事实上,从图 1 中可

见选择的断点并没有将不同决策值的实例完全区分开，产生这种结果的原因是由于单属性离散化算法仅考虑数据在单个属性上的分布，因此算法结果只能取得较好的局部聚类效果.

表 2　单属性离散化后的决策表

| $U$ | $a$ | $b$ | $d$ |
|---|---|---|---|
| 1 | 0 | 1 | 1 |
| 2 | 0 | 1 | 1 |
| 3 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 |
| 5 | 1 | 1 | 1 |
| 6 | 1 | 0 | 0 |
| 7 | 1 | 1 | 0 |

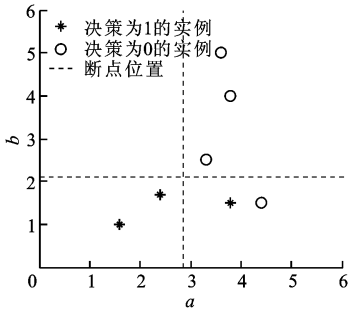


图 1　单属性离散化结果

如图 2 所示，只需要在 $a$ 和 $b$ 的适当位置上选取断点，就可将具有不同决策属性值的实例完全区分开. 由于多属性离散化算法在离散化过程中综合考虑了实例在各个属性上的分布，因此相对于单属性离散化，多属性离散化对断点的选取更合理.

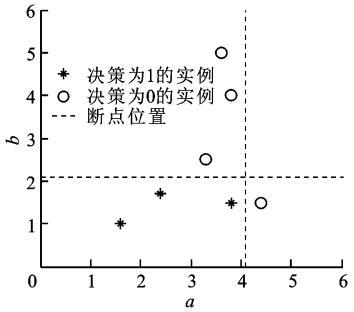


图 2　多属性离散化结果

多属性离散化算法的目标是在数据集的范围中寻求最优断点集，而非单个属性上的最优断点集. 首先，利用 Ameva 算法得到每个属性的最佳断点，然后引入粗糙集理论中的近似分类质量，通过在每一个属性上插入最佳断点，计算由此引起的决策表近似分类质量的变化，选择近似分类质量增量最大的属性作为最佳选择，并在该属性上插入断点；重复以上步骤直到离散化结果满足算法终止条件；最后，剔除已选断点集中的冗余断点，并输出结果.

### 2.1　最佳断点的计算

**定义 3**　候选断点的 Ameva 统计量[8]为

$$A_k = \frac{\chi^2(k)}{k(n-1)} \tag{6}$$

其中

$$\chi^2(k) = N\left(-1 + \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{m_{ij}^2}{m_i m_j}\right)$$

式中：$n$ 表示实例所属的类别总数；$k$ 表示插入断点后将实例分割成的区间数量；$m_i$ 表示所有的实例中属于第 $i$ 类的实例总数量；$m_j$ 表示第 $j$ 个区间中实例的总个数；$m_{ij}$ 表示第 $j$ 个区间中属于 $i$ 类的实例数量；$N$ 表示总的实例数量；$\chi^2$ 表示皮尔逊统计量.

若条件属性 $C$ 与决策属性 $D$ 之间组成列联表，则统计量 $A_k$ 有如下性质：①当属性 $C$ 与 $D$ 之间统计独立时，$A_k$ 具有最小值 0；②$A_k$ 的值越大，属性 $C$ 和 $D$ 之间的相关性就越大，当每一个离散区间 $C_j$ 中实例的决策属性 $D$ 具有相同的值时，$A_k$ 具有最大值；③在 $A_k$ 值的计算中除以 $k(n-1)$ 是为了控制离散化后的区间数量. 因此，使用 $A_k$ 不仅可以最大化 $D$ 和 $C$ 之间的相关性，同时能够使得离散化后的区间数量趋于最少.

### 2.2　插入断点的属性选择

每一个连续属性上都有一个最佳断点，而如何选择属性并在下一步中在该属性上插入断点，将决定连续属性离散化的最终结果. 因此，引入粗糙集理论中的似分类质量，通过在每个属性上预插入最佳断点，以信息系统信息量的变化作为下一步插入断点属性的选择依据.

在每一个属性上试插入该属性的最佳断点，并计算由此引起的决策表近似分类质量的变化，近似分类质量变化越大说明能够获得更多的信息增益，也就是说在该属性上插入断点比在其他属性上插入断点能够获得更好的全局聚类效果.

### 2.3　算法停止法则

算法采用近似分类质量作为判断是否结束的依据. 近似分类质量是用粗糙集理论表示的数据中的信息含量，原始决策表具有最大的近似分类质量. 在信息系统中逐步插入断点的过程中，离散决策表的近似分类质量逐渐提高，当离散决策表的近似分类质量达到原始决策表的近似分类质量时，断点寻找

的算法结束.通过引入近似分类质量,算法保证了信息系统在离散后具有最小的信息损失.

### 2.4 基于 Ameva 的多属性离散化算法

算法在整个数据集范围内寻求最优断点,是一种全局的、至上而下细分的离散化算法.决策表用 $J$ 表示,$U$ 表示全体实例,$B$ 为候选断点集合,$P$ 为已选断点集,$L$ 为实例被 $P$ 划分成的等价信息类,则算法的执行过程如下:

输入 $J$;

输出 $P$ and $L$;

**步骤 1** $P=\varnothing$;$L=\{U\}$;

**步骤 2** 计算每一个属性上候选断点的 $A_k$ 值,具有最大 $A_k$ 者为最佳断点;

**步骤 3** 分别对每一个属性试插入最佳断点,并计算插入断点后决策表近似分类质量的变化;

**步骤 4** 选择具有最大近似分类质量增量的属性,并在该属性上插入最佳断点;

**步骤 5** 将最佳断点加入到 $P$ 中,并从 $B$ 中删除该断点,更新 $P$ 和 $B$;

**步骤 6** 如果此时决策表的近似分类质量达到了原始表的近似分类质量,则转步骤 7,否则转到步骤 2;

**步骤 7** 对于 $P$ 中的断点,若删除后对于决策表的近似分类质量没有影响,则说明该断点是冗余的,且从 $P$ 中剔除;

**步骤 8** 算法结束并输出 $P$ 和 $L$.

## 3 实 验

本文使用 UCI[11] 数据库的 Iris、Wine、Ionos、Breast 数据集对算法进行测试,数据集的基本情况见表 3. Iris 以鸢尾花的特征作为数据来源,数据集包含 150 个样本,分为 3 类,每类 50 个数据,包含 4 个属性;Breast 是乳腺癌数据集,它包含 699 个样本,分为 2 类,共 9 个属性;Wine 是葡萄酒数据集,它包含 178 个样本,分为 3 类,共 13 个属性;Ionos 是电离层数据集,它包含 351 个样本,分为 2 类,共 34 个属性.

表 3 测试样本数据集

| 数据集 | 样本容量/个 | 属性数量/个 | 类别数/个 |
|---|---|---|---|
| 鸢尾花 | 150 | 4 | 3 |
| 乳腺癌 | 699 | 9 | 2 |
| 葡萄酒 | 178 | 13 | 3 |
| 电离层 | 351 | 34 | 2 |

为了验证算法的性能,采用交叉验证对数据进行测试,随机地将样本分成 10 组,并从中挑选 9 组样本作为训练集,剩余的一组样本作为测试集.用本文的算法对训练集做离散化处理,并用获得的划分断点对测试集做离散,然后利用离散化后的训练集对 C45 算法[12]构建分类器,再利用离散后的测试集对训练后的分类器进行测试.实验共进行了 10 次验证,同时记录了 10 次的平均分类错误率和决策树的平均节点数量.表 4、表 5 分别记录了本文算法 AM2 与 ChiMerge 算法、CHI2 算法、ENT_MDLP[13] 算法、Mantaras[14] 算法、Zeta[15] 算法和 Ameva 算法的平均分类错误率和平均节点个数.

表 4 不同算法的平均分类错误率比较

| 数据集 | 平均分类错误率/% | | | | | | |
|---|---|---|---|---|---|---|---|
| | ChiMerge | CHI2 | ENT_MDLP | Mantaras | Zeta | Ameva | AM2 |
| 鸢尾花 | 5.02 | 5.33 | 4.25 | 10.23 | 8.24 | 6.00 | 3.33 |
| 乳腺癌 | 4.92 | 6.80 | 6.37 | 7.79 | 13.05 | 5.42 | 4.38 |
| 葡萄酒 | 7.92 | 6.79 | 7.95 | 7.53 | 6.72 | 6.74 | 5.63 |
| 电离层 | 8.94 | 8.52 | 9.15 | 10.69 | 11.37 | 8.72 | 8.67 |

表 5 不同算法的平均决策树节点数比较

| 数据集 | 平均节点数/个 | | | | | | |
|---|---|---|---|---|---|---|---|
| | ChiMerge | CHI2 | ENT_MDLP | Mantaras | Zeta | Ameva | AM2 |
| 鸢尾花 | 9 | 4 | 7 | 7 | 5 | 4 | 6 |
| 乳腺癌 | 21 | 15 | 21 | 11 | 19 | 23 | 12 |
| 葡萄酒 | 9 | 15 | 15 | 10 | 15 | 9 | 9 |
| 电离层 | 27 | 18 | 11 | 11 | 11 | 13 | 11 |

通过上述测试结果可以看出，对于 4 组数据集，本文提出的算法在分类错误率上有 3 组为最优，其错误率远远低于其他 6 种算法，只有 Ionos 数据集为次优，且非常接近最优结果. 经分析可知，本文算法的分类错误率随着属性数量的增加而提高，Ionos 数据集有 34 个属性，此时的分类错误率达到最高. 从表 5 可以看出，算法在节点的数量上有 2 组数据集是最优的，一组是次优的，因此基于 Ameva 的多属性离散化算法无论是在分类精度还是在决策树节点数上均具有良好的性能.

## 4　结　论

本文对基于统计量的单属性离散化算法进行了研究，并结合粗糙集理论将 Ameva 算法由单属性拓展到多属性，提出了一种新的连续属性离散化算法. 新算法通过计算属性上的断点对决策表近似分类质量的贡献，从全局范围选择断点，能够获得更好的全局聚类效果，在优化划分断点位置的同时避免了有用信息的丢失. 本文给出的算法流程通过实验，验证了算法的有效性.

**参考文献：**


[1] XU E, SHAO Liangshan. A new discretization approach of continuous attributes [C]// Proceedings of the 2010 Asia-Pacific Conference on Wearable Computing Systems: APWC 2010. Piscataway, NJ, USA: IEEE Computer Society, 2010: 136-138.

[2] MIZIANTY M J, KURGAN L A, OGILA M R. Discretization as the enabling technique for the Naive Bayes and semi-Naive Bayes-based classification [J]. Knowledge Engineering Review, 2010, 25(4): 421-449.

[3] KERBER R. ChiMerge: discretization of numeric attributes [C]// Proceedings of Ninth National Conference on Artificial Intelligence. Menlo Park, CA, USA: AAAI Press, 1992: 123-128.

[4] LIU Huan, SETIONO R. Feature selection via discretization [J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9(4): 642-645.

[5] TAY E H, SHEN L. A modified chi2 algorithm for discretization [J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(3): 666-670.

[6] CBAO T S, JYH H H. An extended chi2 algorithm for discretization of real value attributes [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 437-441.

[7] MARCE B. Khiops: a statistical discretization method of continuous attributes [J]. Machine Learning, 2004, 55(1): 53-69.

[8] GONZALEZ A L, CUBEROS F J, VELASCO F. Ameva: an autonomous discretization algorithm [J]. Expert Systems with Applications, 2008, 36(3): 5327-5332.

[9] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001：1-25.

[10] 王国胤. 基于断点辨别力的粗糙集离散化算法 [J]. 重庆邮电大学学报，2010，22(2)：257-261.
WANG Guoyin. Cut's discriminability based continuous attributes discretization in rough set [J]. Journal of Chongqing University of Posts and Telecommunications, 2010, 22(2): 257-261.

[11] MERZ C J, MURPHY P M. UCI repository of machine learning database [EB/OL]. [2003-08-16]. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[12] MITCHELL T M. Machine learning [M]. Beijing, China: China Machine Press, 2003: 52-180.

[13] FAYYAD U, IRANI K. Multi-interval discretization of continuous-valued attributes for classification learning [C]// Thirteenth International Conference on Machine Learning. Bari, Italy: Morgan Kaufmann, 1996: 157-165.

[14] CERQUIDES J, MANTARAS R L. Proposal and empirical comparison of a parallelizable distance based discretization method [C]// Third International Conference on Knowledge Discovery and Data Mining: KDD97. Newport Beach, CA, USA: AAAI Press, 1997: 139-142.

[15] HO K M, SCOTT P D. Zeta: a global method for discretization for continuous variables[C]// Third International Conference on Knowledge Discovery and Data Mining: KDD97. Newport Beach, CA, USA: AAAI Press, 1997: 191-194.


（编辑　管咏梅）