第34卷 第17期 计 算 机 工 程 2008年9月
Vol.34 No.17 Computer Engineering September 2008
·软件技术与数据库· 文章编号：1000—3428(2008)17—0039—03 文献标识码：A 中图分类号：TP18

连续属性离散化的 Imp-Chi2 算法

桑 雨，闫德勤，刘 磊，梁宏霞
(辽宁师范大学计算机信息与技术学院，大连 116029)

摘 要：连续属性离散化是机器学习和数据挖掘领域中的一个重要问题，离散化是否合理决定着表达和提取相关信息的准确性。经过研究Chi2系列算法，提出一种新的基于属性重要性的连续属性离散化方法——Imp-Chi2算法，该算法依据属性重要性程度对属性离散化的顺序进行了合理的调整，能够更准确地对连续属性进行离散化。文章通过C4.5和支持向量机分别对离散化后的结果进行了实验，在实验过程中，提出一种训练集类比例抽取方法，避免了训练集随机抽取的不均匀性。实验结果证明了所提算法的有效性。

关键词：连续属性离散化；Chi2算法；属性重要性；训练集类比例抽取

Imp-Chi2 Algorithm for Discretization of Real Value Attributes

SANG Yu, YAN De-qin, LIU Lei, LIANG Hong-xia
(College of Computer and Information Technology, Liaoning Normal University, Dalian 116029)

【Abstract】 Discretization is an effective technique to deal with continuous attributes for machine learning and data mining. Reasonability of a discretization process is determined by the accuracy of expression and extraction for informations. By analyzing a series of Chi2 algorithm, a new algorithm called Imp-Chi2 algorithm is proposed, which is based on attribute significance. The algorithm reasonably adjusts the sequence of disretization for attributes according to the level of attribute significance, and exactly discretes the real value attributes. The experiments are performed respectively with the results of discreted data by using C4.5 and SVM. In the process of the experiments, a selection method of training set according to class proportion is presented. The method overcomes the bad-distributed situation for random selection of training set. Experimental results show that the presented algorithm is effective.

【Key words】 discretization of real value attributes; Chi2 algorithm; attribute significance; selection of training set according to class proportion

1 概述

在规则提取、特征分类等很多算法中，特别是应用粗糙集理论方法进行数据挖掘的研究和应用中，连续(实值)属性必须进行离散化。连续属性离散化就是把具有实值意义的属性值表示为符号型表示。对连续属性离散化方法的本质要求是最大程度地保持信息表示的意义，减少信息损失。目前，连续属性离散化的方法研究主要有如下几种形式：有监督与无监督，局部与整体，分拆与合并，直接式与增量式。其中最有影响的是有监督形式的基于信息熵思想的相关算法和基于统计学思想的Chi2相关算法。

1992年~2006年期间，学术界相继提出了一系列的Chi2算法[1-4]。在合并标准和离散化步骤上进行了不断的改进，在合并标准上的改进尤其明显。但是，这些算法均忽略了属性离散化顺序上的问题，属性离散化顺序是否合理直接影响着离散化的效果。因此，本文提出了一种基于属性重要性的Chi2算法——Imp-Chi2算法。

2 Chi2系列算法基本概念[3-4]

为便于说明，先介绍Chi2系列算法所涉及到的基本概念。

(1)区间和断点

数据集的初始区间是把每个值作为一个区间，即：区间是属性值的集合。2个相邻的区间根据一个断点进行区分。连续属性离散化实际上就是根据一定的准则，消除断点，合并相邻区间的过程。

(2) χ^2 和 χ^2_α

χ^2 是概率中的统计量，在该离散化算法中需要计算出相临区间的 χ^2 值。χ^2 的计算方法为

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}} \tag{1}$$

其中，k 为决策类别数；A_{ij} 为 i 区间中 j 类样本的个数；$E_{ij}=R_i\times C_j/N$。$R_i=\sum_{j=1}^{k}A_{ij}$ 为 i 区间中样本数，$C_j=\sum_{i=1}^{2}A_{ij}$ 为 j 类样本的个数，$N=\sum_{i=1}^{2}R_i$ 为总样本个数。

χ^2_α 为该离散化方法中的一个参数，用它作为临界值，它由相临2个区间的自由度和显著水平 α 决定。在统计学中，具有 k 个决策类的统计量 χ^2 的渐近分布是自由度为 $k-1$ 的 χ^2 分布，即 $\chi^2_{(k-1)}$ 分布。当给定显著水平 α 时，可确定相应的临界值 χ^2_α，有 $\int_0^{\chi^2_\alpha}f(x)=\alpha$。其中，$f(x)$ 是自由度为 $v=k-1$ 的 χ^2 分布的概率密度函数。事实上，χ^2_α 可通过查表得到。

(3)不一致率

当对象的条件属性相同而决策属性不同时，说明该决策表的分类信息有一定的不一致率，此时，不一致率为

$$Incon_rate=1-\gamma_P$$

基金项目：国家自然科学基金资助项目(60372071)；辽宁省教育厅高等学校科学研究基金资助项目(2004C031)；辽宁师范大学校基金资助项目

作者简介：桑 雨(1982－)，男，硕士研究生，主研方向：数据挖掘和模式识别；闫德勤，教授、博士；刘 磊、梁宏霞，硕士研究生

收稿日期：2007-09-30 **E-mail：**sangyu2008bj@sina.com

其中，γ_P 是近似精度。Imp-Chi2 中利用 $Incon_rate$ 来控制离散化过程中的合并程度和信息丢失。

3 粗糙集

设 $S=(U,A,V,F)$ 为一信息系统。其中，$U=\{x_1,x_2,\cdots,x_n\}$ 是论域；A 是属性集合；V 是属性取值集合；F 是 $U\times A\to V$ 的映射。若 $A=C\cup D$，$C\cap D=\phi$，C 称为条件属性集，D 称为决策属性集，则该信息系统称为决策表。

定义 1 $x,y\in U$，对于 $P\subseteq A$，θ_P 是 U 上的一个等价关系，如果满足 $x\theta_P y\Leftrightarrow(\forall p\in P)(f_p(x)=f_p(y))$，则称 θ_P 是 x，y 的一个不可分辨关系。

定义 2 设 $X\subseteq U$ 为论域的一个子集，$P\subseteq C$，X 的关于 P 的下近似为

$$\underline{P}X=\{x\in U\mid [x]_P\subseteq X\}$$

其中，$[x]_P$ 表示 U 中在等价关系 P 下的等价类元素构成的集合。

定义 3 设 U 为一个论域，P,Q 为 U 上的 2 个等价关系簇，Q 的 P 正域记为 $POS_P(Q)$，定义为

$$POS_P(Q)=\bigcup_{X\in U/Q}\underline{P}(X)$$

定义 4 设 $P\subseteq C$，对于划分 $\{Y_1,Y_2,\cdots,Y_k\}$ 的 P 的近似精度为

$$\gamma_P=\sum_{i=1}^{k}card(\underline{P}Y_i)/card(U)$$

其中，$card(\ )$ 表示集合的基数；γ_P 反映决策表分类的正确程度，描述了关于论域 U 的知识完备程度。

4 Imp-Chi2 算法

连续属性离散化方法是否合理决定着对信息的表达和提取的准确性。在信息系统中，重要的属性对决策划分的影响大，相对于决策属性来说也比较重要。如果先离散化了重要的属性，这样会影响其他属性的合并，信息系统也会过早地出现不一致，所以在对每个属性进行离散化时，希望先合并不重要的属性，这样对其他属性不会产生影响，可以得到更好的离散效果。因此，为了避免 Chi2 系列算法在属性离散化顺序上的不合理性，提出了一种基于属性重要性的 Chi2 算法——Imp-Chi2 算法。该算法依据属性重要性程度对属性离散化的顺序进行了合理的调整，能够更准确地对连续属性进行离散化。Imp-Chi2 算法中的合并标准延续文献[4]中的 $D=(\chi_\alpha^2-\chi^2)/\sqrt{2v}$。

定义 5 设 $S=(U,A,V,F)$ 是一决策表，条件属性子集 $B\subset C$，任意条件属性 $a\in C$ 相对于条件属性集合 B 对决策属性集合 D 依赖程度的属性重要度定义为

$$sgf(a,B,D)=\gamma_{B+\{a\}}-\gamma_B$$

其中，γ_B 是上面提到的近似精度。

Imp-Chi2 算法如下：

Step1 初始化。令显著水平 $\alpha=0.5$。计算信息系统不一致率 $Incon_rate$。

Step2 对每个属性将数据排序并根据式(1)计算所有相临区间的 χ^2 值，通过查表找出与 χ^2 对应的 χ_α^2 值，再计算差异 D。

Step3 合并。

```
while(尚有可合并的断点)
 {寻找 D 最大的断点进行合并;
 if(Incon_rate 增大)
  {撤消合并;   goto Step4;}
 else   goto Step2;}
```

Step4

```
if ( α 已是最后一级)
   {退出程序，离散化完毕;}
else { α0 = α ;
       对 α 降级;
       goto Step2;}
```

Step5 对每个属性进行离散化。

对当前信息系统，计算每个属性相对于条件属性集的重要度，找到重要度最小的属性 i(设 n 为条件属性个数)

```
k=n;
while (k>0)
 {对属性 i
   {计算差异 D;
    α = α0 ;
    标志 flag=0;
    while(flag==0)
      {while(尚有可合并的断点)
         {寻找 D 最大的断点进行合并;
          if(Incon_rate 增大)
              {撤消合并; flag=1; break;}
          else 更新差异 D;   }
       if( α 已是最后一级)   break;
       else {对 α 降级;  更新差异 D;}
       }
    }
  k--;
```

对当前信息系统，计算每个属性相对于条件属性集的重要度，找到重要度最小的且没有合并过的属性 i }

5 训练集类比例抽取方法

为更好地选择训练集，笔者提出了类比例抽取的方法。训练集类比例抽取的思想是：从数据集的每个类中均匀地抽取出样本，每个类中抽取出的样本数量是有一定比例的，即：在每个类中抽取出 $L=t_i\cdot(T/N)$ 个样本，再把从各个类中抽取出的样本组成训练集。这样，抽取出的训练集每个类中都有一定比例的样本数。其中，T 为训练集个数，$1\le i\le c$；c 为决策类别数；N 为样本总数；t_i 为 i 类的样本数。

这种方法避免了训练集随机抽取的不均匀性。例如，某数据集有 3 个类，那么随机抽取出的训练集很有可能只抽到 2 个类(相对不好的情况)或者第 3 类抽到的数量很少，这样训练后得到的模型不好，预测精度也不会高。所以，选用训练集类比例抽取方法来选取训练集可以得到更加稳固的训练模型，提高测试精度。

6 实验与结果

笔者从 UCI 机器学习数据库中选取了 6 个数据集(见表 1)，每个数据都是一致的。

表 1 数据信息

数据集	连续属性	离散属性	类别数	样本数
Iris	4	0	3	150
Breast	9	0	2	683
Wine	13	0	3	178
Auto	5	2	3	392
Bupa	6	0	2	345
Machine	7	0	8	209

6 个数据集均用本文所提出的 Imp-Chi2 算法(简称 Imp)和文献[4]的方法(简称 Ext)进行了离散化，对离散化后的数据应用 C4.5 方法构造决策树，随机选取 80%作为训练集，其余 20%作为测试集。对统计平均正确识别率、错误识别率和拒

绝识别率以及平均决策树节点个数(node)和提取出规则的平均个数(rule)进行对比(见表 2)。

表 2　C4.5 识别实验结果

数据集	正确识别率		错误识别率		拒绝识别率		结点个数		规则个数	
	Ext	Imp	Ext	Imp	Ext	Imp	Ext	Imp	Ext	Imp
Iris	0.916 7	0.916 7	0.031 6	0.031 6	0.051 7	0.051 7	20.85	20.85	14.35	14.35
Breast	0.925 5	0.926 6	0.040 9	0.042 3	0.033 6	0.031 1	89.10	86.90	57.50	54.40
Wine	0.802 8	0.918 1	0.093 1	0.044 4	0.104 1	0.037 5	62.80	40.25	36.00	22.40
Auto	0.789 9	0.797 5	0.085 4	0.072 8	0.124 7	0.129 7	125.80	127.85	89.95	92.35
Bupa	0.437 9	0.465 9	0.205 7	0.215 9	0.356 4	0.038 2	288.80	238.80	177.00	180.30
Machine	0.773 8	0.773 8	0.216 7	0.216 7	0.009 5	0.009 5	64.15	64.15	42.45	42.45

同时，使用 SVM 对离散数据分别用“一对一”、“一对多”和“DAG”3 种多类分类方法进行分类[5]，随机选取 80%作为训练集，其余 20%作为测试集。模型类型选为 C-SVC，核函数类型选为 RBF 函数，惩罚因子 C 搜索范围：[1, 100]，核函数参数 γ 取 0.5。对 2 种算法所统计的预测精度(用 acc 表示)和支持向量个数(用 svs 表示)进行了对比(见表 3)。

表 3　SVM 分类预测结果

数据集	一对一				一对多				DAG			
	Ext		Imp		Ext		Imp		Ext		Imp	
	acc	svs	acc	svs	acc	svs	acc	svs	acc	svs	acc	svs
Iris	0.933	36	0.967	18	0.933	83	0.967	21	0.933	36	0.967	18
Breast	0.971	91	0.978	110	0.971	91	0.978	110	0.971	91	0.978	110
Wine	0.972	37	1.000	29	0.972	45	1.000	33	0.972	45	1.000	29
Auto	0.696	148	0.797	139	0.696	154	0.785	149	0.696	192	0.810	139
Bupa	0.681	180	0.710	219	0.681	181	0.710	219	0.681	176	0.710	219
Machine	0.690	101	0.809	74	0.667	107	0.786	106	0.714	72	0.809	74

由于核函数依赖于输入样本向量的内积，大的属性值容易导致计算复杂，训练时间较长，为避免上述情况发生，将属性值进行归一化：

$$\bar{x}_i = 2\frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} - 1$$

归一化后的属性值 $x_i \in [-1,+1]$。训练样本与测试样本采用相同的归一化方法。

从表 2 中可以看出，对 Iris 和 Machine 数据集来说，新的算法 Imp-Chi2 与文献[4]中的方法具有相同的识别率。对 Breast, Auto 和 Bupa 数据集而言，新算法的平均正确识别率均有所上升，Breast 数据集的平均决策树结点个数和平均决策树规则个数均有所下降，这 2 点正是离散化方法期望得到的结果，同时，Bupa 数据集的平均决策树结点个数也有所下降。对 Wine 数据集来说，Imp-Chi2 算法正确识别率得到显著提升，平均决策树节点个数和提取出规则的平均个数显著下降，充分显示了 Imp-Chi2 算法的优势。

同时，从表 3 中也可以看出，对于 6 个数据集而言，与文献[4]中方法相比，在 3 种分类方法下，Imp-Chi2 算法的预测精度均有显著增加，数据集 Iris, Wine, Auto 和 Machine 的支持向量个数均有所减少，表现出 Imp-Chi2 算法的有效性。

对于表 2、表 3 中的实验，训练集的选取都是随机选取的，这样有时会导致选取的不均匀，从而得不到稳固的模型。用训练集类比例抽取方法与训练集随机抽取作了以下实验对比(见表 4 和表 5)。其中，Imp(类比例)为采用训练集类比例抽取，Imp 为采用训练集随机抽取。

从实验结果可以看到，对训练集类比例抽取方法而言，表 4 中 6 个数据集的平均正确识别率均有明显的增加。Iris, Breast, Auto 和 Bupa 数据集的平均决策树结点个数和平均决策树规则个数多一些，Machine 和 Wine 数据集的平均决策树结点个数和平均决策树规则个数少一些。在表 5 中，对数据集 Auto 和 Machine 的预测精度来说，训练集类比例抽取方法比训练集随机抽取要好一些，对于其余的数据集，2 种训练集抽取方法的预测精度相同。对 6 个数据集的支持向量来说，2 种训练集抽取方法实验所得到的支持向量个数相近，可以说明训练集类比例抽取方法是有效的。

表 4　C4.5 识别实验结果

数据集	正确识别率		错误识别率		拒绝识别率		结点个数		规则个数	
	Imp	Imp (类比例)	Imp	Imp (类比例)	Imp	Imp (类比例)	Imp	Imp (类比例)	Imp	Imp (类比例)
Iris	0.916 7	1.000 0	0.031 6	0.000 0	0.0517	0.00	20.85	23	14.35	17
Breast	0.926 6	1.000 0	0.042 3	0.000 0	0.0311	0.00	86.90	94	54.40	62
Wine	0.918 1	1.000 0	0.044 4	0.000 0	0.0375	0.00	40.25	24	22.40	18
Auto	0.797 5	1.000 0	0.072 8	0.000 0	0.1297	0.00	127.85	132	92.35	101
Bupa	0.465 9	0.938 4	0.215 9	0.061 6	0.0382	0.00	238.80	257	180.30	200
Machine	0.773 8	0.842 9	0.216 7	0.157 1	0.0095	0.00	64.15	52	42.45	38

表 5　SVM 分类预测结果

数据集	一对一				一对多				DAG			
	Imp		Imp(类比例)		Imp		Imp(类比例)		Imp		Imp(类比例)	
	acc	svs	acc	svs	acc	svs	acc	svs	acc	svs	acc	svs
Iris	0.967	18	0.967	22	0.967	21	0.967	28	0.967	18	0.967	21
Breast	0.978	110	0.978	98	0.978	110	0.978	98	0.978	110	0.978	98
Wine	1.000	29	1.000	26	1.000	33	1.000	42	1.000	29	1.000	26
Auto	0.797	139	0.810	151	0.785	149	0.785	142	0.810	139	0.823	135
Bupa	0.710	219	0.710	214	0.710	219	0.710	214	0.710	219	0.710	214
Machine	0.809	74	0.889	60	0.786	106	0.889	62	0.809	74	0.889	60

从表 2~表 5 的实验结果分析可以说明本文所提方法的优势和有效性。

7　结束语

本文提出的 Imp-Chi2 算法能够更合理更准确地对连续属性进行离散化，解决了 Chi2 系列算法在属性离散化顺序上存在的不合理性。同时，使用 C4.5 和 SVM 对离散化后的结果进行了实验。实验结果证明了本文所提算法的有效性，得到了很好的效果。

参考文献

[1] Kerber R. ChiMerge: Discretization of Numeric Attributes[C]//Proc. of the 9th National Conference on Artificial Intelligence. [S. l.]: AAAI Press, 1992: 123-128.

[2] Liu Huan, Setiono R. Feature Selection via Discretization[J]. IEEE Trans. on Knowledge and Data Eng., 1997, 9(4): 642-645.

[3] Tay E H, Shen Lixiang. A Modified Chi2 Algorithm for Discretization[J]. IEEE Trans. on Knowledge and Data Eng., 2002, 14(3): 666-670.

[4] Chao-Ton Su, Jyh-Hwa Hsu. An Extended Chi2 Algorithm for Discretization of Real Value Attributes[J]. IEEE Trans. on Knowledge and Data Eng., 2005, 17(3): 437-441.

[5] Hsu C W. A Comparison of Methods for Multiclass Support Vector Machines[J]. IEEE Trans. on Neural Networks, 2002, 13(2): 415-425.