文章编号：1000-8608(2012)03-0443-05

基于改进χ^2统计的数据离散化算法

桑 雨[1]， 李克秋*[1]， 闫德勤[2]

（1.大连理工大学 计算机科学与技术学院，辽宁 大连 116024；
2.辽宁师范大学 计算机与信息技术学院，辽宁 大连 116029）

摘要： 在基于χ^2统计独立性的离散化算法中，自由度与期望频数的选取直接影响χ^2计算的准确性，从而影响离散化的性能.为此，提出了一种基于改进χ^2统计的数据离散化算法，提高了基于统计独立性离散化算法的质量.首先，分析了χ^2函数中自由度选取的不足，给出了自由度选取的修正方案；其次，根据数据类分布等特点，提出了期望频数的改进方案，克服了不同数据集赋予相同期望频数的缺陷，提高了χ^2计算的准确性.实验结果表明，改进的方法显著提高了C4.5决策树与Naive贝叶斯分类器的学习精度.

关键词： 离散化；数据挖掘；χ^2统计
中图分类号： TP18 **文献标志码：** A

0 引 言

随着数据库中信息量的增加以及信息管理水平的不断提高，涌现了各种类型的数据来描述客观世界.在应用机器学习从数据中提取知识时，涉及的数据通常包括离散值(如男、女)和连续值(如身高、温度等).然而，大多数的数据挖掘、归纳学习等算法仅仅适用于使用离散化方法描述的样本，如C4.5[1]和AQ算法[2]等.因此，连续属性必须进行离散化，其实质是分割连续属性的值域，转化成若干个有意义的区间，简化数据，提高分类器的学习精度.

离散化算法的类型有[3]考虑类信息的有监督类型和不考虑类信息的无监督类型；考虑整体样本的全局型和考虑部分样本的局部型；相邻区间合并的自底向上型(bottom-up)和区间分割的自顶向下型(top-down).EQW和EQF[3]是实现简单且计算消耗低的自顶向下无监督离散化算法.著名的自顶向下有监督离散化算法包括基于信息熵理论的算法，如Ent-MDLP[4]；基于类属性相互依赖的算法，如CACC[5].Ent-MDLP通过定义信息熵标准来最小化模型总的信息量，同时利用MDLP来决定合适的离散区间数.CACC是目前最新的基于类-属性相互依赖的离散化算法，它提出了一个启发式断点选择标准，考虑了所有样本的分布信息，并且避免了过拟合现象，产生了理想的离散化方案.著名的自底向上有监督离散化算法包括基于统计学理论的Chi2-based相关算法[6~9]，如ChiMerge[6]和Extended Chi2[9]等.它们首先初始化区间，采用χ^2统计来判断当前相邻区间是否被合并，并且通过不一致衡量标准来判断离散化进程是否结束.

基于χ^2统计的方法是目前最有效的离散化算法之一.自由度与期望频数的选取直接影响χ^2计算的准确性，从而影响离散化的性能.本文提出一种基于改进χ^2统计的数据离散化算法，该算法考虑相邻区间数对自由度的影响.此外，对于没有在相邻区间中出现的类，期望频数均取一个预先给定的常数，忽视了自身的内在信息对期望频数的影响，导致计算χ^2不准确，区间合并顺序不合理，从而降低了学习精度.因此，本文给出自由度与期望频数的合理改进方案.

1 基础知识

1.1 粗糙集[10]

设$S=(U,A,V,F)$是一个信息系统，其中

收稿日期： 2010-09-18； **修回日期：** 2012-03-28.
基金项目： 教育部新世纪优秀人才支持计划资助项目(NCET-07-0132).
作者简介： 桑 雨(1982-)，男，博士生；李克秋*(1971-)，男，博士，教授，博士生导师，E-mail：keqiu@dlut.edu.cn.

$U=\{x_1,x_2,\cdots,x_n\}$ 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. A 由条件属性集合 C 与一个决策类属性 d 来决定，即 $A=C\cup d$，$C\cap d=\varnothing$，则此信息系统被定义为决策表.

对于 $x,y\in U$，$P\subseteq A$ 是 U 上的一个子集(等价关系)，如果满足 $xPy\Leftrightarrow(\forall p\in P)(f_p(x)=f_p(y))$，则 x 和 y 是在等价关系 P 下所构成等价类集合中的元素.

定义 1 假设论域 U 的一个子集为 X，条件属性集合 C 的一个子集为 P，则 X 关于 P 的下近似被定义为

$$P_-X=\{x\in U\mid [x]_P\subseteq X\}$$

其中 $[x]_P$ 是 P 所产生等价类的元素构成的集合.

定义 2 等价关系 P 关于决策类属性所划分的等价类 $\{Y_1,Y_2,\cdots,Y_k\}$ 的一致性水平为

$$\gamma_P=\frac{\sum_{i=1}^{k}card(P_-Y_i)}{card(U)} \tag{1}$$

其中 $card(\cdot)$ 是集合的基数.

1.2 χ^2 统计

离散化任务要求训练集包含 N 个样本，每个样本属于 k 个类中的其中一类，且包含 m 个连续属性(条件属性). 基于 χ^2 统计的自底向上离散化算法的实质是在所有相邻区间对中决定哪一对相邻区间首先被选择合并.

χ^2 统计可评价被离散区间与类属性之间的独立性. 在离散化过程中，需要计算所有相邻区间的 χ^2 来判断当前哪对区间先被合并，计算方法如下：

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

式中：k 为数据集总的决策类别数；A_{ij} 为 i 区间 j 类样本数；$E_{ij}=R_i\times C_j/M$，为 A_{ij} 的期望频数，其中 $R_i=\sum_{j=1}^{k}A_{ij}$ 为 i 区间样本数，$C_j=\sum_{i=1}^{2}A_{ij}$ 为相邻两区间中 j 类样本数，$M=\sum_{i=1}^{2}R_i$ 为相邻两区间的样本总数. 如果 $C_j=0$，则 $E_{ij}=0.1$.

2 改进 χ^2 统计的离散化算法

有效的离散化标准(区间合并标准)可以产生好的离散化结果. 在基于 χ^2 统计的离散化算法中，χ^2 统计的合理性直接影响离散化的性能. 然而，χ^2 统计中在自由度的选取上仅仅考虑了相邻区间的类别数，忽视了相邻区间数的作用；另外，对于没有在相邻区间中出现的类，其期望频数取一个预先给定的常数，忽视了数据类分布对期望频数取值的影响. 这些缺陷导致计算 χ^2 不准确，区间合并顺序不合理，从而降低学习精度. 基于上面两点不足，本文提出一种基于改进 χ^2 统计的数据离散化算法，该算法考虑了相邻区间数对自由度的影响，并依据数据类分布给出了合理的期望频数，能够合理准确地进行离散化. 下面，具体分析这两点不足并且提出有效的改进方案.

2.1 χ^2 分布中自由度选取的不足及改进

在 Chi2 算法[7]中，χ^2 分布的自由度选取为 $\upsilon=k-1$，k 为数据总的决策类别数. 改进的 Chi2 算法[8]认为自由度的选择应该根据划分断点两边区间的类别数来确定，即 $\upsilon=k'-1$，k' 为相邻区间对中的类别数，$2\leqslant k'\leqslant k$.

一般来说，χ^2 分布的随机变量为 $W=Z_1^2+Z_2^2+\cdots+Z_n^2$，其中 Z_i 服从标准正态分布，$i=1,2,\cdots,n$. 也就是说，W 服从自由度为 $n-1$ 的 χ^2 分布，即 χ^2_{n-1} 分布. 这样，式(2)中的 χ^2 相当于 χ^2 分布的随机变量，由 $2k'$ 项的 Z_i^2 加和获得，即服从自由度为 $2k'-1$ 的 χ^2 分布. 在自由度的选取上，χ^2 统计仅仅考虑了相邻区间的类别数，忽视了相邻区间数的作用；换句话说，自由度选取不仅仅与相邻区间的类别数有关，还与相邻区间数有关. 因此，应该选取 $\upsilon=2k'-1$ 作为 χ^2 统计显著性检验的自由度.

2.2 χ^2 分布中期望频数 E_{ij} 取值的不足及改进

在式(2)中，如果 $C_j=0$，则 $E_{ij}=0.1$. 也就是说，如果相邻区间的类别数小于总的类别数($k'<k$)，则对于没有在相邻区间中出现的类，其期望频数 E_{ij} 均取一个预先给定的值 0.1. 然而，这忽视了数据类分布对 E_{ij} 取值的影响，导致计算 χ^2 不准确以及不合理的区间合并顺序.

假设存在两对相邻区间，其中一个区间对的类别数大于另一对的，且与数据集总类数不等. 然而，如果类别较多的区间对的类分布较均匀，而类别较少的区间对的类分布不均匀，考虑公平性，如

果区间对类别较多，则适当降低 E_{ij} 取值；如果区间对类别较少，则适当增加 E_{ij} 取值. 基于上面分析，针对数据本身的特点，本文启发式地选取 $(2k-v)/2k$ 作为 E_{ij} 取值的重要部分. 然而，当数据集总的类别数为3时，$(2k-v)/2k$ 中的 v 是常量，因此，不能区分出各对相邻区间 χ^2 函数中 E_{ij} 的差异. 本文考虑了相邻区间对的自由度与区间大小的相关关系，即自由度越大，区间样本数越多；自由度越小，区间样本数越少，所以，选取 $(N-M)/N$ 作为 E_{ij} 取值的另一部分. 总之，如果相邻两区间的自由度较大，则使 E_{ij} 按较小比例增加；如果相邻两区间的自由度较小，则使 E_{ij} 按较大比例增加. 基于上面的分析，有以下改进方案：

如果 $k'<k$，并且 $C_j=0$，则有

$$E_{ij}=2\cdot[(2k-v)/2k]\cdot[(N-M)/N] \tag{3}$$

其中 N 为数据集总的样本数，$v=2k'-1$（以改进的自由度为标准），$2\leqslant k'\leqslant k$，$i\in\{1,2\}$，$1<j\leqslant k$.

注意：式(3)中 $[(2k-v)/2k]\cdot[(N-M)/N]$ 前面乘以2有以下原因.

首先引进 χ^2_α 的概念：χ^2_α 为Chi2-based离散化算法中的重要参量，由相邻区间 χ^2 分布的自由度和显著水平 α 决定. 如果给定 α，则临界值 χ^2_α 相应地被确定. 事实上，χ^2_α 可以通过查表得到. 由于Chi2-based算法中 α 的初始值被设为0.5，如果 $v=2k'-1=3$（k' 最小取2），$\chi^2_\alpha=2.365\,5$，并且在Chi2-based算法中，必须使得 $\chi^2\leqslant\chi^2_\alpha$，这样，选取 E_{ij} 值的上限为小于2.365 5的适当的值2.

从上面的分析中可以看到，式(3)可以完整地反映出 E_{ij} 取值在 χ^2 统计中的合理性，并很好地解决了 χ^2 统计应用在Chi2-based算法中的缺陷.

2.3　算法描述

本文所提出的基于改进 χ^2 统计的数据离散化算法基于的是Chi2-based算法的框架，分为两个阶段进行离散化：第一阶段考虑整体属性进行区间合并，算法通过不一致衡量标准自动地进行离散化，当被离散数据的不一致率超过原始数据不一致率时，算法停止；第二阶段对每个属性进行离散化，使得离散化更加精确. 注意：新算法的区间合并标准为差异 $D=(\chi^2_\alpha-\chi^2)/\sqrt{2v}$，与Extended Chi2算法[9]相似，不同的是，本文提出的 χ^2_α 和 χ^2 采用的是第2章中改进的算法.

基于改进 χ^2 统计的离散化算法描述如下.

第一阶段：

步骤1　设置显著水平 $\alpha=0.5$，根据式(1)计算数据的一致性水平 γ_c.

步骤2　升序排序每个连续属性的值，计算所有相邻区间改进后的 χ^2 以及差异 D.

步骤3　考虑整体连续属性，选择合适的相邻区间进行合并

```
while(存在相邻区间对)
{ 合并最大 D 值的相邻区间；
  if(γc 下降)
  { 取消合并；跳到步骤 4；}
  else 返回步骤 2；
}
```

步骤4　if α存在下一级别 then

```
{     α0 = α;
      将 α降一级别，返回步骤 2；
}
else  停止离散化
```

第二阶段：对单个连续属性离散化，精炼区间

```
for {每个连续属性}
{ α = α0；
  设置标志 sign=0；
  while (sign==0)
  {  while (存在相邻区间对)
     { 合并最大 D 值的相邻区间；
       if(γc 下降)
       {取消合并；sign=1；break；}
       else 更新差异 D；
     }
  }
```

如果 α 不能降级，停止离散化；否则，将 α 降级，更新差异 D.

```
}
```

下面，对新算法的时间复杂度做具体分析. 每个连续属性值排序的时间复杂度为 $O(N\log N)$；对于本文提出的算法，对 χ^2 统计量做了两处改进：一是自由度的改进；原始自由度 $v=k'-1$ 与改进自由度 $v=2k'-1$ 都是通过相邻两区间中的

类别数 k' 决定的，然而，求得相邻区间类别数的时间复杂度是不发生变化的，因此，自由度的改进没有影响求得差异 D 所需时间的变化. 二是 χ^2 中 E_{ij} 取值的改进；对于 Extended Chi2 算法，计算 χ^2 的时间复杂度为 $O(2kN)$，改进 E_{ij} 取值后，计算 χ^2 的时间复杂度为 $O(2kN)+O(M)=O(2kN)$，这里 $M<N$. 综上，在对 χ^2 统计量改进后，不会影响求得差异 D 所需时间的变化. 由于所提出算法的框架相似于 Extended Chi2 的框架，新算法的时间复杂度仍为 $O(KmN\log N)$，其中 m 为连续属性个数，K 为算法的增量步数.

3 性能评价

在实验中，采用了 UCI 机器学习数据库[11]中的 9 个数据集(见表 1)来评价本文所提算法的性能. 数据集均是数据挖掘等实验所常用的. 将所提出的基于改进 χ^2 统计的离散化算法与下列 4 种算法进行了比较.

(1) EQF：经典的无监督离散化算法[3]；

(2) Ent-MDLP：基于熵的离散化算法[4]；

(3) Ext-Chi2：最先进的自底向上离散化算法[9]；

(4) CACC：最先进的自顶向下离散化算法[5].

表 1 数据集描述

Tab. 1 Description of datasets

数据集	连续属性	离散属性	类别数	样本数
Iris	4	0	3	150
Auto	5	2	3	392
Breast	9	0	2	683
Ionosphere	34	0	2	351
Pima	8	0	2	768
Glass	9	0	6	214
Vehicle	18	0	4	846
Vowel	10	3	6	990
Page-blocks	10	0	5	5 473

9 个数据集全部通过上述离散化算法进行离散化，在 VC++6.0 环境下实现. 将离散后的数据应用 C4.5 方法构造决策树，并采用 Naive 贝叶斯分类器进行分类预测，使用 Weka 数据挖掘工具[12]进行分类预测，采用 10 折交叉验证的方法[13]对平均学习精度统计进行对比(见表 2 和 3).

表 2 C4.5 分类预测结果

Tab. 2 Classification and prediction results by C4.5 %

数据集	离散化算法				
	本文算法	CACC	Ext-Chi2	Ent-MDLP	EQF
Iris	94.6	92.8	93.8	93.4	92.8
Auto	83.7	78.3	81.2	79.6	72.7
Breast	96.7	96.2	96.4	96.1	93.5
Ionosphere	91.9	90.6	92.8	86.5	82.7
Pima	77.6	74.9	75.3	70.0	67.4
Glass	78.6	77.5	70.1	73.5	63.8
Vehicle	72.3	67.7	70.6	67.3	65.2
Vowel	97.3	97.7	97.9	95.5	95.1
Page-blocks	96.9	95.8	95.8	95.2	94.3

表 3 Naive 贝叶斯分类预测结果

Tab. 3 Classification and prediction results by Naive Bayes %

数据集	离散化方法				
	本文算法	CACC	Ext-Chi2	Ent-MDLP	EQF
Iris	96.3	92.5	95.4	94.1	91.7
Auto	79.5	80.0	75.4	76.3	74.1
Breast	97.9	97.2	96.5	94.8	94.2
Ionosphere	94.8	92.1	93.5	94.3	92.9
Pima	73.5	68.1	67.5	70.5	69.8
Glass	76.2	67.8	70.4	59.7	63.6
Vehicle	68.9	67.2	67.9	65.8	67.7
Vowel	96.7	94.1	94.9	93.2	91.2
Page-blocks	96.8	92.3	91.7	93.5	88.7

由表 2 可以看出，在 9 个数据集上，本文算法的平均分类精度有所提高. 由于 EQF、Ent-MDLP 和 CACC 均没有考虑离散化过程中的数据信息丢失情况，与本文算法和 Ext-Chi2 算法相比，这 3 种算法有较低的分类精度.

由表 3 可以看出，在正确识别率方面，本文算法的平均学习精度是最高的，可见，当 χ^2 统计量改善后数据的平均学习精度显著提高，充分显示了本文所提算法的有效性.

4 结 语

基于概率统计理论的 Chi2 系列算法为连续属性离散化算法的研究提供了新的思路. 本文分析了 Chi2 系列算法中 χ^2 统计量的不足，并提出了合理的改进方案，获得了期望的离散化结果，提高了分类器的学习精度.

参考文献：

[1] QUINLAN J R. **C4.5: Programs for Machine Learning** [M]. San Mateo: Morgan Kaufmann, 1993

[2] MICHALSKI R S, MOZETIC I, HONG Ja-rong, *et al*. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains [C] // **Proceedings of Fifth National Conference on Artificial Intelligence**. Pennsylvania: AAAI Press, 1986: 1041-1045

[3] DOUGHERTY J, KOHAVI R, SAHAMI M. Supervised and unsupervised discretization of continuous feature [C] // **Proceedings of 12th International Conference of Machine Learning**. San Mateo: Morgan Kaufmann, 1995: 194-202

[4] FAYYAD U, IRANI K. Multi-interval discretization of continuous-valued attributes for classification learning [C] // **Proceedings of Thirteenth International Joint Conference on Artificial Intelligence**. San Mateo: Morgan Kaufmann, 1993: 1022-1027

[5] TSAI C J, LEE C I, YANG W P. A discretization algorithm based on class-attributes contingency coefficient [J]. **Information Sciences**, 2008, **178**(17): 714-731

[6] KERBER R. ChiMerge: discretization of numeric attributes [C] // **Proceedings of Ninth National Conference on Artificial Intelligence**. San Jose: AAAI Press, 1992: 123-128

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

[8] TAY E H, SHEN L. A modified Chi2 algorithm for discretization [J]. **IEEE Transactions on Knowledge and Data Engineering**, 2002, **14**(3): 666-670

[9] SU C T, HSU J H. An extended Chi2 algorithm for discretization of real value attributes [J]. **IEEE Transactions on Knowledge and Data Engineering**, 2005, **17**(3): 437-441

[10] PAWLAK Z. Rough sets [J]. **International Journal of Computer and Information Sciences**, 1982, **11**(5): 341-356

[11] HETTICH S, BAY S D. The UCI KDD archive [DB/OL]. [2010-08-25]. http://kdd.ics.uci.edu/, 1999

[12] PENTAHO. Weka 3: data mining software in Java [EB/OL]. [2010-08-25]. http://www.cs.waikato.ac.nz/ml/weka, 2007

[13] WEISS S M, KULIKOWSKI C A. Computer systems that learn: classification and prediction methods from statistics, neural nets [M] // **Machine Learning and Expert Systems**. San Mateo: Morgan Kaufmann, 1990

A data discretization algorithm based on improved chi-square statistic

SANG Yu[1], LI Ke-qiu[*1], YAN De-qin[2]

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;
2. School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract: The selection of degree of freedom and expected frequency directly affects the accuracy of chi-square calculation in discretization algorithms based on chi-square statistical independence. This will affect the performance of discretization. A data discretization method based on improved chi-square statistic is proposed. It improves the quality of discretization algorithm based on statistical independence. Firstly, the deficiency of the selection of degree of freedom in chi-square function is analyzed, and a modified scheme for selection of degree of freedom is given. Secondly, an improved scheme for expected frequency is proposed according to data class distribution, which overcomes the deficiency that different datasets have the same expected frequency. This improves the accuracy of chi-square calculation. The experimental results show that the improved algorithm improves the learning accuracy of C4.5 decision tree and Naive Bayes classifier.

Key words: discretization; data mining; chi-square statistic