连续特征离散化方法综述

汪庆，张巍，刘鹏

上海财经大学信息管理与工程学院，上海，200439
wq_811@yahoo.com.cn

摘要：离散特征在数据挖掘的过程中具有重要的作用，如何将实际数据集中的连续特征最好地离散化是一个 NP-hard 问题。本文介绍了离散化方法的不同角度的分类、离散化过程中用到的术语及一般离散化的过程。同时，还按照离散化方法有监督和无监督的分类方法体系，介绍了几种有代表性的离散化方法。提出需根据学习环境选择合适的离散化方法，将关联分析中连续特征离散化作为以后的研究方向。

关键词：连续特征，离散特征，离散化

1 引言

数据集的特征按照其取值可以分为连续特征和离散特征。连续特征也称为定量特征，通常用间隔尺度和比例尺度来衡量，有较多甚至无穷的数值表达, 其值取自某个连续的区间, 表示了对象的某种可测性质,例如人的身高、年龄, 商品的价格、空气温度、物体长度等等。离散特征也称定性特征，一般以名义尺度或有序尺度定义，是指以文本型数据表达的对象特征, 如人的性别、学历特征, 商品的用途(食品、服装) 等, 此类特征的值域只限定于较少的取值[1-2]。连续特征的取值允许被排序，可进行算术运算；离散特征的取值有时允许被排序，但是其不能进行算术运算[3]。

在机器学习和数据挖掘中，已经发展了处理离散型数据的很多算法，如决策树、关联规则及基于粗糙集理论的许多方法，而这些算法对于连续型数据却不适用；而有些算法即使能处理连续型数据，挖掘和学习也没有处理离散型数据有用和有效。但是在实际数据库中，往往不只存在着离散型数据，也存在着大量连续型数据。这样就有必要将连续特征离散化，使得特征可以适用于各种算法。特征的离散化处理就是把连续特征转化为离散特征，它是数据预处理的一个重要过程，直接关系到挖掘和学习的效果[3-4]。

将连续特征离散化，再将离散化的结果应用于算法有很多好处。（1）离散化结果将会减少给定连续特征值的个数，减小系统对存储空间的实际需求。（2）离散特征相对于连续特征来说更接近于知识层面的表示。（3）通过离散化，数据被规约和简化，对于使用者和专家来说，离散化的数据都更易于理解，使用和解释。（4）离散化处理使得算法的学习更为准确和迅速[5]。（5）一系列算法只能应用于离散型数据，使得离散化处理成为必要，而离散化又使很多算法的应用范围扩展了[4, 6, 7]。但最优离散化问题已经被证明是一个NP-hard问题。

离散化的方法有很多，本文接下来第 2 节介绍了离散化方法的分类体系、术语及离散化过程，第 3 节选取了目前比较有代表性的几种离散化方法进行了详细介绍以及一些改进的离散化方法，第 4 节提出了要根据学习环境和用户需要选择合适的离散化方法，并以关联分析中的离散化为例，指出在关联分析中离散化方法选择需要注意的问题，最后是全文的结束语，并将关联分析中的连续特征离散化作为以后的研究方向。

2 现状及离散化过程

2.1 分类

离散化方法依据不同的需求沿着不同的主线发展至今，目前已存在很多不同离散化方法的分类体系。不同的分类体系强调离散化方法间的区别的不同方面[3]。主要的分类体系有有监督的和无监督的、动态的和静态的、全局的和局部的、分裂式的（从上至下）和合并式的（从下至上）、单变量的和多变量的以及直接的和增量式的。

根据离散化方法是否使用数据集的类信息，离散化方法可以分为有监督的和无监督的。有监督的离散化方法使用类信息，而无监督的离散化方法不使用类信息。有监督的离散化方法又分为建立在错误率基础上的、建立在熵值基础上的或者建立在统计信息基础上的[3, 5, 8]。早期的等宽、等频的离散化方法是无监督方法的典型代表，连续的区间根据使用者给定的宽度或频数划分成小的区间。无监督的方法的缺陷在于它对分布不均匀的数据不适用，对异常

点比较敏感[9]。为了克服无监督的离散化方法的这些缺陷，使用类信息来进行离散化的有监督的离散化方法逐渐发展起来。

离散化方法也常以动态或静态的分类方法来区分。动态的离散化方法就是在建立分类模型的同时对连续特征进行离散化，如有名的C4.5[10]。静态的离散化方法就是在进行分类之前完成离散化处理。在Doughertyet al. [5]文中有动态和静态离散化方法的详细对比。

根据离散化过程是否是针对整个训练数据空间的，离散化方法又可分为全局的和局部的。全局的离散化方法使用所有的实例，而局部的离散化方法只是用一部分的实例。

离散化方法还可分为从上至下的和从下至上的，也可称为分裂式的和合并式的。分裂的离散化方法起始的分裂点列表是空的，通过离散化过程逐渐往列表中加入分裂点，而合并的离散化方法则是将所有的连续值都看作可能的分裂点，再逐渐合并相邻区域的值形成区间。

单变量的离散化方法是指一次只对数据集的一个特征进行离散化，而多变量的离散化是同时考虑数据集的多个特征及其相互关联关系进行离散化，需要考虑更多的因素，算法更加复杂[11]。

另外一种离散化方法的分类是直接式的和增量式的。直接式的离散化方法就是根据额外给定的参数（离散化所需得到的区间数等）一次性形成所有的分裂点，而增量式的离散化方法是根据某个准则逐渐的将离散化结果进行改进，直到满足准则的停止条件为止[3, 6-7]。

Liu Huan 和 Hussian [6]在其文中提出了离散化方法的一个层次框架，主要根据合并分裂及有无监督的分类体系来划分离散化方法，比较全面的概括了离散化方法的分类及各类别中的代表离散化方法，具体参见文献Liu and Hussian（2002）[6]。层次框架见图 1。

2.2 术语

数据集由实例（Instance）组成，而每一个实例又通过一组特征（Feature）来描述。连续特征离散化就是采取各种方法将连续的区间划分为小的区间，并将这连续的小区间与离散的值关联起来。

离散化中最常用到的一个术语是断点(Cut Point)。断点就是小区间的划分点，区间的一部分数据小于等于断点值，另一部分数据则大于断点值。选取断点的方法不同，使得产生了不同的离散化方法。

最终离散得到的区间的个数称为离散的区间数（Arity）。较大的离散区间数会使得数据的可理解性变差，但较低的离散区间数又会对预测的准确性造成影响。各种离散化方法的离散区间数的确定也不相同，有的方法是直接人工确定，做为外部变量输入，而另外一些离散化

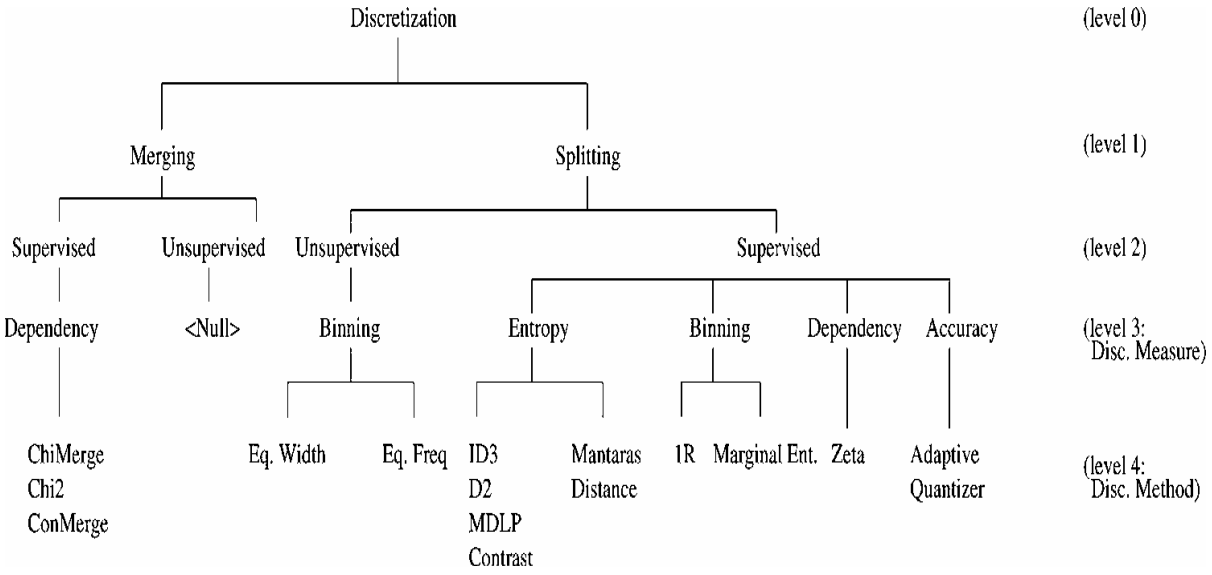

图 1 离散化方法层次框架[6]

方法则是不需要外部输入离散区间数，在算法过程中由某些准则来停止离散过程，从而得到合适的离散区间数。

2.3 离散化处理的一般过程

对连续特征进行离散化处理，一般经过以下步骤：

（1）对此特征进行排序。特别是对于大数据集，排序算法的选择要有助于节省时间，提高效率，减少离散化的整个过程的时间开支及复杂度。

（2）选择某个点作为候选断点，用所选取的具体的离散化方法的尺度进行衡量此候选断点是否满足要求。

（3）若候选断点满足离散化的衡量尺度，则对数据集进行分裂或合并，再选择下一个候选断点，重复步骤（2）（3）。

（4）当离散算法存在停止准则时，如果满足停止准则，则不再进行离散化过程，从而得到最终的离散结果。

2.4 离散化结果的评价

不同的离散化方法会产生不同的离散化结果。优良的离散化，应使划分尽可能简约，又尽可能多的保留由样本数据代表的对象的固有特性[12]。

离散化结果的好坏可以从以下几方面来考虑：(1)区间的个数。这也是对模型简洁性的要求。理论上来说，离散得到的区间数越少越好，便于理解，但区间数的减少另一方面也会导致数据的可理解性变差; (2) 离散化所导致的不一致性。离散化之后数据的不一致性不能比离散化之前更高。这一点是对模型一致性的要求。（3）预测准确性。即对模型准确性的要求。这一点通常通过交叉检验模式建立分类树来衡量[6]。

3.离散化方法及其改进

离散化主要是用于分类或者关联分析中的特征离散化。通常来说，离散化结果的好坏依赖于采用的离散化算法，以及需要考虑的其他特征。

一般现在的离散化算法都是考虑单个特征进行离散化，而不考虑对多个特征的关系进行分析同时进行离散化[13]。目前已存在很多对连续特征进行离散化的方法，这些方法既可以是单一算法的，也可以是合成的。单一的算法在完成离散化的过程中不会用到其他的离散化方法，而合成的方法在离散化的过程中则同时用到多种离散化算法的思想，是建立在多种单一算法的基础之上的。

本文接下来按照有监督的和无监督的离散化方法分类体系介绍几种比较典型的离散化的方法。在离散化之前，假定待离散的数据已按升序排列，根据离散化需要排序的要求。

3.1 无监督的方法

3.1.1 分箱法

分箱法包括等宽分箱法和等频分箱法，它们是基本的离散化算法[7]。

分箱的方法是基于箱的指定个数自顶向下的分裂技术，在离散化的过程中不使用类信息，属于无监督的离散化方法。等宽分箱法，就是根据箱的个数得出固定的宽度，使得分到每个箱中的数据的宽度是相等的。等频分箱法是使得分到每个箱中的数据的个数是相同的。在等宽或等频划分后，可用箱中的中位数或者平均值替换箱中的每个值，实现特征的离散化。

这两种方法简单，易于操作，但都需要人为地规定划分区间的个数这个参数。同时，使用等宽法的缺点在于它对异常点比较敏感[13]，倾向于不均匀地把实例分布到各个箱中，有些箱中包括许多实例，而另外一些箱中又一个实例都没有。这样会严重地损坏特征建立好的决策结构的能力。而等频的方法虽然避免了上述问题的产生，却可能将具有相同类标签的相同特征值分入不同的箱中以满足箱中数据的固定个数的条件[8]。

对于等宽法对异常点敏感的问题， 其中一种可用的方法是离散化前首先设定某个阈值将异常数据移除。针对等频法易将具有相同类标签相同特征值的实例分入不同箱中的问题，可用的解决方法是先用等频法将特征值进行分箱，然后对各个相邻分箱的边界值进行调整，使得相同的值可以被分入同一个箱中[6]。

3.1.2 根据直观划分离散化

对于实际数据的离散化，用户希望离散后得到的区间是相对一致，易于阅读，看上去直观或者自然的区间[7]。如对于年薪的划分，用户更希望看到的结果是[150000，60000]，而不是由下文将提到的基于熵或者聚类算法得到的[15235.9，60872.43]这样的区间。根据直观划分的离散化方法使用 3-4-5 规则可以将特征的数值区间划分为相对一致和自然的区间。

一般地，3-4-5 规则根据最高有效位的取值范围，递归逐层地将给定的数据区域划分为 3、4 或者 5 个相对等宽的区间。使用此规则进行离散化的具体过程可参见Han Jiawei and

Kamber（2004）[7]。现实世界中的数据常包含特别大的正的和负的离群值，基于最小和最大数据值得自顶向下的离散化方法可能导致扭曲的结果。根据直观划分离散化的方法先将数据集中的数值得大多数的数值区间（如，第 5 个百分位数到第 95 个百分位数）挑出来作为顶层离散化区间按照 3-4-5 规则进行离散化，再将超出顶层离散化区间的特别高的和特别低的值用类似的方法单独处理成不同的区间。

这种方法的优点主要在于它适用于实际数据的简单且有实际意义的离散化，并且特别考虑了离群点的离散化。

3.1.3 基于聚类分析的离散化

基于聚类分析的离散化方法也是一种无监督的离散化方法。此种方法包含两个步骤，首先是将某特征的值用聚类算法（如K-means算法）通过考虑特征值的分布以及数据点的邻近性，划分成簇或者组。然后是将聚类得到的簇进行再处理，可分为自顶向下的分裂策略和自底向上的合并策略。分裂策略是将每一个初始簇进一步分裂为若干子簇，合并策略则是通过反复地对邻近簇进行合并[3]。聚类分析的离散化方法也需要用户指定簇的个数，从而决定离散产生的区间数。

现阶段，无监督的方法还比较少，在没有类信息的情况下，要得到好的离散化结果比较困难，并且离散化的结果也比较难衡量。但是实际数据集在多数情况下又是没有类标签的，可以考虑先使用聚类算法人为地为数据集添加类标签，然后再用添加了类标签的数据集进行离散化指导离散化过程。

3.2 有监督的方法

3.2.1 1R 方法

1R[14]是一种使用分箱的有监督的方法。它把连续的区间分成小的区间，然后再使用类标签对小区间的边界进行调整。每个区间至少包含 6 个实例，除了最后一个区间外，最后一个区间包含所有未被列入其他区间的实例。从第一个实例开始，把前 6 个实例列入第一区间，如果下一个实例与此区间中大多数实例的类标签相同，则把此实例加入区间中，再判定下一个实例按照前述操作能否加入刚才的区间中，否则形成下一个含 6 个实例的新的区间，对下一个实例重复类似的操作，直至结束。然后把区间中的大多数实例的共同类标签作为此区间的类标签，如果相邻区间经过此操作后有相同的类标签，则应把这两个相邻区间合并。

1R离散化方法也是分箱方法，操作仍然比较方便，但又不再需要用户人为指定箱的个数，也克服了无监督的分箱方法的不使用类信息的缺陷，并且能避免把具有相同特征值相同类标签的实例分入不同的小区间中[6]。

3.2.2 基于熵的离散化方法

由于建立决策树时用熵[15]来分裂连续特征的准则在实际中运行得很好，考虑将这种思想扩展到更通常的离散化中，通过反复地分裂小区间直到满足停止的条件。由此产生了基于熵的离散化方法。熵是最常用的离散化度量之一。基于熵的离散化方法使用类分布信息计算和确定分裂点，是一种有监督的、自顶向下的分裂技术[7]。

ID3[16]和C4.5[10]是两种常用的使用熵的度量准则来建立决策树的算法。ID3 通过贪心算法搜寻给定数据区间内的具有熵值最小的数据点作为断点。该方法将区间内的每一个数值作为候选断点，计算其熵值，然后从中选出具有最小熵值的数据点作为断点，将区间一分为二，然后再对得到的区间递归地应用以上方法进行离散化。停止准则是当得到的每个区间中的类标签都是一样时，即停止离散化过程。这个准则也可以根据需要适当放宽要求。它的缺点在于使用的停止准则的原则化合理性是有待考虑的。分类和离散化毕竟是两个不能完全等同的处理过程。而使用ID3 和C4.5 的方法来离散化完全是照搬利用这两种算法建立决策树的思想。

在上法的基础上，又产生了D2[9]和MDLP（minimum description length principle，最小描述长度准则）[17]的基于熵的离散化方法。D2 法与ID3 法不同之处在于，ID3 法是动态的，在建树的过程中进行离散化，而D2 法则不是。除此外，D2 法与ID3 法没有明显的不同，D2 法也是计算数据集中每个取值的熵，选取熵值最小的点作为端点，将区间一分为二，再对得到的每一个区间递归地二分。离散化过程在满足D2 法的停止准则时再停止。这些停止准则包括待分裂的区间中含的实例数少于 14 个、或者得到的区间数大于等于 8 个等等。

MDLP的思想是假设断点是类的分界，依此得到许多小的区间，每个区间中的实例的类标签都是一样的，然后再应用MDLP准则衡量类的分界点中哪些是符合要求可以作为端点，哪些不是端点需要将相邻区间进行合并。由此选出必要的断点，对整个数据集进行离散化处理。这种离散化方法的停止准则相对于D2 更完善，并且选出的断点是区分类的点[6]。

3.2.3 基于卡方的离散化方法

前面提到的离散化方法均是采用自顶向下的分裂策略，即先把整个数据集当作一个区间，再逐步选出端点对大的区间进行分裂得到小的区间，而基于卡方的离散化方法是采用自底向上的策略，首先将数据取值范围内的所有数据值列为一个单独的区间，再递归地找出最佳邻近可合并的区间，然后合并他们，进而形成较大的区间[7]。在判定最佳邻近可合并的区间时，会用到卡方统计量来检测两个对象间的相关度。

最常用的基于卡方的离散化方法是ChiMerge方法[18]，它是一种自动化的离散化算法[1]。它的过程如下：首先将数值特征的每个不同值看作一个区间，对每对相邻区间计算卡方统计量，将其与由给定的置信水平确定的阈值进行比较，高于阈值则把相邻区间进行合并，因为高的卡方统计量表示这两个相邻区间具有相似的类分布，而具有相似类分步的区间应当进行合并形成为一个区间。合并的过程递归地进行，直至计算得到的卡方统计量不再大于阈值，也就是说，找不到相邻的区间可以进行合并，则离散化过程终止，得到最终的离散化结果。

应用卡方统计量检验两个对象是否相关时，需要人为设定置信水平参数，由统计学知识算出一个与计算量相比较的阈值。对于置信水平的设置要合理，过高会导致过分离散化，过低又会导致离散化不足。对此Liu and Setiono（1997）[19]等人作了改进,并形成Chi2 算法。它不再使用固定显著性水平，而是使显著性水平逐步下降，这样可以使得在满足不一致率准则的前提下，更多的区间被合并。Chi2 算法分为两个阶段进行离散化，并通过检验系统的不一致率，确定是否终止.。

用不一致率检验离散化程度的 Chi2 算法，优于显著性水平固定设定的 Chimerge 算法,但该算法也有弱点，主要有： 1) 需设定不一致率的下限值,为此需经多次试验,工作量大，且难以优选；2) 不一致率并不能全面反映分类特性，而这些特性体现了对样本数据离散化程度的要求；3) 有多个变量(特征) 需要离散化时，Chi2 算法没有考虑离散化顺序对结果的影响。

除上面介绍的几种离散化方法之外，还有很多其他的离散化的方法，包括在前述方法基础上针对其缺陷进行改进的离散化方法，以距离度量等其它思想作为指导进行离散化的方法，以及在一种离散化方法中揉合应用多种单一的离散化方法以有效结合各种单一离散化方法的优点克服某些缺点的方法。由于篇幅限制，本文并未详细介绍所有的离散化方法。

4. 根据学习环境选择离散化方法

需要注意的是，虽然已有很多离散化方法，但是没有一种离散化方法对任何数据集以及任何算法都是有效的，也没有一种离散化方法一定比其他方法产生更好的离散化结果。因为离散化本身就是一个NP-hard 问题，所以在使用时一定要根据数据集的特点和学习环境以及使用者个人的偏好理解等选择合适的离散化方法，以取得尽可能好的离散化效果。如，决策树学习容易受到碎片问题的影响，所以离散化时更偏好得到较少的离散化区间；决策规则希望离散化得到的区间中的实例的类标签是唯一的；关联规则重视特征间的相关性，所以在离散化时不能对各个特征进行单一的离散化[3]。

以关联规则分析为例。在进行关联规则分析时，需要重视特征间的联系，而不能对数据集的各个特征进行单一的离散化。对各个特征进行孤立的离散化会破坏原始数据集的特征内在相关性，从而导致推导出不完善的关联规则，影响关联规则分析的正确性和有效性。所以关联规则分析中的连续特征离散化应考虑选择多变量的离散化方法。

并且，关联规则分析中采用的数据集在某些情况下是没有类信息的，无法使用有监督的离散化方法来进行连续特征的离散化。故在离散化时应考虑采用无监督的离散化方法，而现阶段无监督的离散化方法的发展还比较有限，主要是等宽等频离散化方法及基于聚类的离散化方法，故关联规则分析中的离散化问题将是一个很值得研究的方面，有效地解决关联分析中连续特征的离散化问题也将有效地促进关联规则分析。-

此外，针对其他的算法和学习环境，也应相应的分析离散化的要求，从而选择合适的离散化算法。

5. 结束语

连续特征离散化过程在数据的预处理中发挥着重要的作用。本文介绍了离散化问题的产生，是为了解决某些算法对于连续数据学习不够高效有用或者只适用于离散型数据的不足，而离散化过程也将带来很多好处。文中也阐述了离散化方法的分类体系、离散化过程中常用的术语、及离散化的一般过程，并根据离散化方法有监督和无监督的分类体系选择了现有的

一些比较常用的离散化的方法，如等宽离散法、等频离散法、基于熵的离散化、基于卡方的离散化等方法进行介绍，阐述了各法的基本思想，及其停止准则等，对有些方法的不足之处也做了简要说明。由于文章篇幅限制，本为并未全面介绍现阶段已有的所有的离散化方法。

注意到没有任何离散化算法可以适用于任何环境下，在实际应用时，需要根据数据集的特点和学习环境等选择合适的离散化方法，而关联规则分析中的离散化既需要考虑各特征间的内在联系，又要考虑在没有类信息的情况下对数据集进行有效的离散化，离散化处理的要求比较高，将作为以后的研究方向进行深入研究。

参考文献

[1] Mehmed Kantardzic， 2003. Data Mining: Concepts, Models, Methods, and Algorithms, IEEE press, pp:19-22,54-58

[2] 张永，丁洪昌,2007。连续特征离散化的 MaxDiff 方法。计算机工程与应用，2007，43（19）

[3] Ying Yang and Xindong Wu,2007 。Discretization Methods Simon, H.A. 1981. The Sciences of the Artificial, 2nd edn. Cambridge, MA: MIT Press.

[4] 刘业政，焦宁等,2007。连续特征离散化算法比较研究。计算机应用研究，2007 年 9 月第 24 卷第 9 期。

[5] Dougherty, J.,Kohavi, R., and Sahami, M. 1995. Supervised and unsupervised discretization of continuous features. In Proc. Twelfth International Conference on Machine Learning. Los Altos, CA: Morgan Kaufmann, pp. 194–202.

[6] HUAN LIU, RARHAD HUSSAIN, CHEW LIM TAN, MANORANJAN DASH, 2002. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery,6,393-423,2002. 2002 Kluwer Academic Publishers. Manufactured in The Netherlands.

[7] Jiawei Han, Micheline Kamber 著，范明，孟小峰等译，2004。数据挖掘概念与技术,第 2 版，机械工业出版社：47-60。

[8] Ian H. Witten, Eibe Frank, 2005.数据挖掘 实用机器学习技术, 英文版第 2 版。机械工业出版社：396-305.

[9] Catlett, J. 1991. On changing continuous attributes into ordered discrete attributes. In Proc. Fifth European Working Session on Learning. Berlin: Springer-Verlag, pp. 164–177.

[10] Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann.

[11] Bay, S. D. 2000. UCI ， KDD archive. Department of Information and Computer Sciences, University of California, Irvine.http://kdd.ics.uci.edu/.

[12] ROBERT S. Analyzing discretizations of continuous attributes given a monotonic discrimination function [J] . Intelligent Data Analysis , 1997 , 1 :157 179.

[13] Pang-Ning Tan and Michael Steinbach，Vipin Kumar，2006。Introduction to Data Mining.英文版，人民邮电出版社：42-45.

[14] Holte, R.C. 1993. Very simple classification rules perform well on most commonly used datasets. Machine Learning, 11:63–90.

[15] Shannon, C. and Weaver, W. 1949. The Mathmatical Theory of Information. Urbana: University of Illinois Press.

[16] Quinlan, J.R. 1986. Induction of decision trees. Machine Learning,1:81-106

[17] Fayyad, U. and Irani, K. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In Proc. Thirteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 1022–1027.

[18] Kerber, R. 1992. Chimerge: Discretization of numeric attributes. In Proc. AAAI-92, Ninth National Confrerence Articial Intelligence. AAAI Press/The MIT Press, pp. 123–128.

[19] HUAN Liu , RUDY Setiono,1997. Feature selection via dis2 cretization[J] . IEEE Transactions on Knowledge and Data Engineering , 1997 ,9 (4) : 642 645