离散技术的调查：监督学习中的分类学和经验分析

文章 in IEEE交易在知识和数据工程上·2013年2月 DOI: 10.1109/TKDE.2012.35

离散技术的调查：监督学习中的分类学和经验分析

salvador garc ´a，Juli ´luengo，jos ´

摘要 - 策划是许多知识发现和数据挖掘任务中使用的必不可少的预处理技术。它的主要目标是通过将分类值与间隔关联，从而将定量数据转换为定性数据，将一组连续属性转换为离散的属性。以这种方式，可以在连续数据上应用符号数据挖掘算法，并简化信息的表示，从而使其更加简洁和特定。文献提供了许多离散化的建议，并可以找到将它们分类为分类法的一些尝试。但是，在以前的论文中，在财产的定义方面缺乏共识，尚未建立正式的分类，这可能会使从业者感到困惑。此外，仅考虑了一小部分离散剂，而许多其他方法也没有引起人们的注意。为了减轻这些问题，本文从理论和经验的角度提供了文献中提出的离散方法的调查。从理论的角度来看，我们根据先前研究中指出的主要属性开发了一种分类法，统一了符号，并包括所有已知方法。从经验上讲，我们在涉及最具代表性和最新的分散剂，不同类型的分类器和大量数据集的监督分类中进行了一项实验研究。通过非参数统计检验，已经验证了其以准确性，间隔数量和不一致性的测量结果的结果。此外，一组离散剂被突出显示为表现最好的人。

索引条款-DISC 重新启动，延续 UOU属性，决策树，分类学，数据prepro 静止，数据挖掘，分类。

1简介

K现在提取和数据挖掘（DM）是要在包含与实际应用程序相关的数据的不同数据库中执行的重要方法[1]，[2]。这两个过程通常都需要一些以前的任务，例如问题理解，数据理解或数据预处理，以便成功地应用DM算法到真实数据[3]，[4]。数据预处理[5]是DMFILD中的一个关键研究主题，其中包括几个数据转换，清洁和降低数据的过程。作为基本数据减少技术之一，疾病近年来引起了研究的越来越多[6]，并且已成为DM中最广泛使用的预处理技术之一。

离散化过程将定量数据转换为定性数据，即具有有限数量的间隔数的数值属性，将数值属性转换为离散或名义属性，获得了一个非重叠分区

- S. García is with the Department of Computer Science, University of Jaén, 23071, Jaén, Spain. E-mail: sglopez@ujaen.es
- J. Luengo is with the Department of Civil Engineering, LSI, University of Burgos, 09006, Burgos, Spain. E-mail: jluengo@ubu.es
- J.A. Sáez, V. López and F. Herrera are with the Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, 18071 Granada, Spain. E-mails: smja@decsai.ugr.es, vlopez@decsai.ugr.es, herrera@decsai.ugr.es

连续域。然后建立每个间隔与数值离散值之间的关联。在实践中，可以将离散化视为一种数据减少方法，因为它将数据从大量数字值映射到大大降低离散值的子集。一旦执行离散化，可以将数据视为任何归纳或扣除DM过程中的标称数据。许多现有的DM算法旨在仅使用名义属性在分类数据中学习，而实际应用程序通常涉及连续功能。在使用此类算法之前，必须将这些数值特征离散化。

在监督的学习和特定的分类中，该调查的主题是，我们可以按以下方式定义分解。假设一个数据集由N示例和C目标类组成，则分散算法将在此数据集中在m exteet Intervals $D \;= \{\; d_0, d_1$ [v14} [v14} [v15] , $(d_1, d_2$, $(d_1, d_2$, $(d_1, d_2$, $(d_1, d_2$ {v12 { {v15 , $(d_1, d_2$ {v12 , $(d_1, d_2$ {v1 , $(d_1, d_2$ {v15 {其中d_0是最小值，d_m是最大值和$d_i \;<\; d_{i+i}$，对于$i \;=0$, 1 , . . . , $m\;-1$。这样一个离散结果D称为属性A}的分散方案A和A}}}}}} $P\;=$}}} { A。

在数据上使用离散化的必要性可能是由几个因素引起的。许多DM算法主要针对处理名义属性[7]，[6]，[8]，甚至可能仅处理离散属性。例如，被认为是DM前十种的十种方法中的三种[9]需要嵌入或外部

Digital Object Indentifier 10.1109/TKDE.2012.35

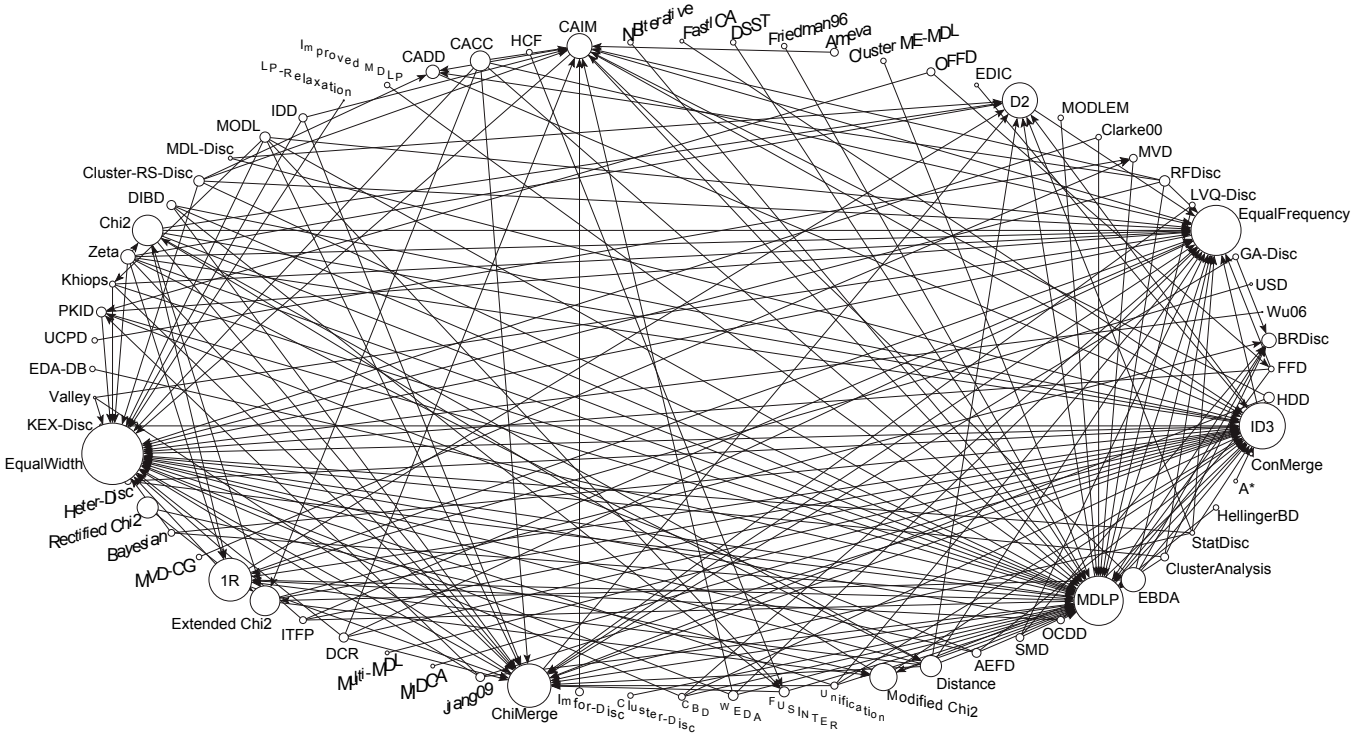


图1：离散器的比较网络。后来，这些方法将在表1中定义。

数据离散化：C4.5 [10]，Apriori [11]和Naive Bayes [12]，[13]。即使使用能够处理连续数据的算法，学习也降低了有效和有效[14]，[5]，[4]。由离散化得出的其他优点是数据的减少和简化，从而使学习速度更快并产生更准确，紧凑和更短的结果；数据中可能存在的噪声减少。对于研究人员和从业者来说，离散属性更容易理解，使用和解释[6]。然而，任何离散化过程通常会导致信息丢失，从而最大程度地将此类信息丢失成为离散器的主要目标。

获得最佳离散化是NP完整的[15]。文献中可以找到大量的离散技术。显然，在处理具体的问题或数据集时，选择权限的选择将调节后验学习任务在准确性，模型的简单性等方面的成功。 Sets [21]，[22]等。为了提供离散剂的分类，例如单变量/多变量，监督/无监督，上下/bottoum-up，Global/lobal/local/local/local，static/dynamic等。所有这些标准都是已经提出的分类法的基础，在本文中将深入阐述。在每种情况下，识别最佳疾病的识别是一项非常困难的任务

出去，但是考虑一组代表性的学习者和离散者，进行详尽的实验可以帮助决定最佳选择。

关于离散技术的一些评论可以在文献[7]，[6]，[23]，[8]中找到。但是，这些方法的特征尚未完全研究，许多离散剂，甚至是经典的特征，也没有提及，并且用于分类的符号也不统一。例如，在[7]中，静态/动态区别与[6]中的静态/动态区别不同，并且全球/局部支持通常与单变量/多变量属性相混淆[24]，[25]，[26]。随后的论文包括一条符号或另一种符号，具体取决于其引用的初始离散研究：[7]，[24]或[6]。

尽管文学丰富，除了使用统一符号对离散化的完全分类外，可以观察到，很少有尝试将它们比较它们。通过这种方式，通常将提出的算法与完整的离散剂家族的子集进行比较，在大多数研究中，未进行严格的经验分析。此外，近年来已经提出了许多新方法，并且对于众所周知的调查中审查的离散化因素[7]，[6]，它们并未引起人们的注意。图1说明了一个比较净工作，其中每个节点对应于离散化算法，并且两个节点之间的有向顶点表明启动节点的算法已将其与末端节点的算法进行了比较。箭头的分量始终是从最新方法到最古老的方法，但并不影响结果。这

节点的大小与输入和输出顶点的数量相关。我们可以看到，大多数离散器都由小节点表示，并且该图远非完整，这促使了本文。比较最多的技术是相等的，相等的频率，MDLP [16]，ID3 [10]，Chimerge [17]，1R [27]，D2 [28]和Chi2 [18]。

这些原因激发了本文的全球目的，可以将其分为三个主要目标：

- 提出基于离散方法中观察到的主要特性的完整分类学。分类法将使我们能够从理论的角度选择离散器来表征他们的优势和缺点。
- 从获得的间隔数量和数据的不一致水平方面，对最具代表性和最新的离散剂进行了实证研究。
- 最后，使用两个指标将一组代表性DM模型的最佳离散器联系起来，以衡量预测性分类成功。

实验研究将包括基于非参数测试的统计分析。我们将进行总共30个离散剂的实验； 6属于懒惰，规则，决策树和贝叶斯学习家庭的分类方法；和40个数据集。鉴于手头的数据，实验评估并不对应于每个离散器的最佳参数的详尽搜索。然后，其主要重点是将最佳性能离散器的一部分与每个经典分类器相关联，并使用对其进行一般性配置。

本文的组织如下。第2节中提供了有关离散化的相关和高级工作。第3节介绍了审查的离散者，其职权和提议的分类法。第4节描述了实验框架，研究了实证研究中所产生的结果，并提出了对其进行讨论。第5节总结了论文。最后，我们必须指出，本文具有关联的网站http://sci2s.ugr.es/discretization，该网站收集了有关实验中涉及的离散剂的其他信息，例如实现和详细的实验结果。

2相关和高级沃尔 k

目前，改善和分析离散化的研究很普遍，需求量很高。根据DM任务，离散化是一种获得希望结果的有前途的技术，这是其与其他方法和问题的关系。本节简要摘要与离散化有关的主题从理论和实用的角度密切相关，并描述了过去几年已经研究的其他作品和未来趋势。

- *Discretization Specific Analysis:* susmaga提出了一种基于连续属性和粗糙集的双向差异的分析方法措施[29]。他强调，他的分析方法可用于检测差异化的冗余以及可以在不降低性能的情况下删除的切口集。此外，它可以应用于改善现有的离散化方法。
- *Optimal Multisplitting:* Elomaa和Rousu介绍了一些基本特性，用于在监督的单变量离散化中使用一些经典的评估功能。他们分析了熵，信息增益，增益比，训练集误差，GINI指数和归一化距离度量，得出的结论是，它们适合用于属性的最佳多幅度[30]。他们还开发了一种用于执行此多幅度过程的最佳算法，并设计了两种技术[31]，[32]来加快速度。
- *Discretization of Continuous Labels:* 在连续监督的学习（回归问题）转换为名义监督学习（分类问题）中，已经使用了两个可能的处理。第一个只是使用回归树算法，例如CART [33]。第二个包括将离散化应用于静态[34]或动态方式[35]的分配。
- *Fuzzy Discretization:* 围绕语言术语的定义进行了广泛的研究，该语言术语将域归因于模糊区域的定义[36]。模糊离散化的特征在于，与属性值相对应的成员价值，组或间隔数和属性的特征，这与仅考虑间隔数字的清晰差异不同[37]。
- *Cost-Sensitive Discretization:* 基于成本的离散化的目的是考虑犯错的成本，而不仅仅是最大程度地减少错误的总和[38]。它与不平衡或成本敏感的分类问题有关[39]，[40]。
- *Semi-Supervised Discretization:* [41]中已经设计了第一次尝试在半监督分类概率中解散数据的尝试，表明它渐近等同于监督的APARACH。

本节中提到的研究不超出本调查的范围。我们指出，本文的主要观点是对文献中发现的离散方法进行广泛概述，并对最相关的离散器进行详尽的实验比较，而没有考虑上述和先进的因素，例如上述或从经典监督的分类中得出的问题。

3个离散化：背景和技术

本节提出了离散化的分类法和用于构建它的标准。首先，在第3.1条中，将定义的主要特征

将概述分类法的类别。然后，在第3.2小节中，我们列举了文献中提出的离散方法，我们将通过使用其完整和缩写的名称以及相关的参考来考虑。最后，我们提出分类学。

3.1离散方法的常见特性

本节为讨论下一个小节中介绍的离散器的讨论提供了一个框架。讨论的问题包括分类学结构中涉及的几个属性，因为它们是离散器运作的独有的。尽管不参与分类法，但将提出其他较少关键的问题，例如参数属性或停止条件。最后，还将指出一些标准以比较离散方法。

3.1.1 Main Characteristics of a Discretizer

在[6]，[7]，[8]中，已经描述了各种轴，以对离散方法进行分类。我们在本节中审查和解释它们，强调它们之间发现的主要方面和关系并统一符号。提出的分类法将基于这些特征：

- *Static vs. Dynamic:* 这种特征是指离散器与学习者相关的时刻和独立性。当学习者正在构建模型时，动态的离散器会起作用，因此他们只能访问嵌入学习者本身中的部分信息（本地属性，请参阅后面），从而与相关的学习者产生紧凑而准确的结果。否则，在学习任务之前将进行静态裁判率，并且它独立于学习算法[6]。几乎所有已知的离散器都是静态的，因为在处理数值数据时，大多数动态离散器实际上是DM算法的子部分或阶段[42]。众所周知的动态技术的一些示例是ID3离散器[10]和ITFP [43]。

- *Univariate vs. Multivariate:* 多变量技术，也称为2D离散化[44]，同时考虑所有属性来定义最初的切口集或完全决定最佳切割点。在研究与其他属性的相互作用时，他们还可以离散一个属性，从而利用高级关系。相比之下，一旦建立了属性之间的顺序，并且在每个属性中，单变量离散器一次仅与单个属性一起工作，并且每个属性中所得的离散方案在后期阶段保持不变。最近出现了兴趣，因为它们在演绎学习中非常有影感[45]，[46]，并且在复杂的分类问题中非常有影响，在这些问题中存在多个属性之间的高度相互作用，而单变量离散器可能会避免[47]，[47]，[48]。

- *Supervised vs. Unsupervised:* 无监督的分散者不考虑班级标签，而超级属性的标签则不考虑。后者考虑类属性的方式取决于输入属性和类标签之间的相互作用，以及用于确定最佳切割点（熵，相互依存等）的启发式测度。文献中提出的大多数分散剂都是监督的，从理论上讲，使用类信息应自动确定每个属性的最佳数量。如果没有得到裁判，则并不意味着它不能在监督任务上应用。但是，只有在监督的DM问题上才能应用监督的疾病。代表性的无监督离散器是相等的宽度和相等的频率[49]，PKID和FFD [12]和MVD [45]。

- *Splitting vs. Merging:* 这是指用于创建或定义新间隔的过程。分裂方法在所有可能的边界点之间建立一个切点，并将域分为两个间隔。相比之下，合并方法从预先确定的分区开始，然后删除候选切口点以混合两个相邻的间隔。这些支撑物分别与*Top-Down*和*Bottom- up*分别高度相关（在下一部分中进行了解释）。它们背后的想法非常相似，只是根据层次的离散化构建，自上而下或自下而上的离散器假设该过程是渐进的（后面描述的）。实际上，可能会有一个离散的人，其操作一次是基于分裂或合并一个以上的间隔[50]，[51]。同样，由于可以在运行时间合并[52]，[53]的合并时，可以将某些离散器视为*hybrid*。

- *Global vs. Local:* 为了做出决定，离散器可以要求属性中的所有可用数据，也可以仅使用部分信息。当仅根据本地信息做出分区决策时，据说离散器是本地的。广泛使用的本地技术的示例是MDLP [16]和ID3 [10]。很少有离散器是本地的，除了基于自上而下的分区和所有动态技术外。在自上而下的过程中，某些算法遵循划分和争议方案，当发现分裂时，数据会递归分配，从而限制了对部分数据的访问。关于动态离散器，它们在DM算法的内部操作中找到了剪切点，因此他们永远无法访问完整的数据集。

- *Direct vs. Incremental:* 直接离散器同时将范围分为k间隔，需要一个附加标准来确定k的值。它们不仅包括一步离散的方法，还包括在操作中执行多个阶段的离散化合物，在每个步骤中选择多个切口。相比之下，incremen-

TAL方法从简单的离散化开始，并通过改进过程，需要一个其他标准才能知道何时停止它。在每个步骤中，他们发现最佳的候选边界被用作切口，然后将其余的决策做出相应。增量分离器也称为分层离散器[23]。两种类型的离散剂在文献中都很普遍，尽管增量和监督者之间通常存在更明确的关系。

- *Evaluation Measure:* 这是离散器用来比较两个候选方案的指标，并确定更适合使用的候选方案。我们考虑了评估措施的五个主要家庭：

  - *Information:*该家族包括*entropy*作为离散化最常用的评估度量（MDLP [16]，ID3 [10]，Fusinter [54]）和其他派生的信息理论测量指标，例如*Gini index* [55]。 - *Statistical:*统计评估涉及属性之间的依赖/相关性（Zeta [56]，Chimerge [17]，Chi2 [18]），概率和贝叶斯属性[19]（MODL [20]），相互依赖[57]，相互依存[57]，偶然性共同的[58]等[58]等，odsection [58]等。通过使用粗略的设定度量和属性[21]（例如下部和上部近似值，类可分离性等）来离散化方案 - *Wrapper:*该集合包括依赖于每次评估的分类器提供的错误的方法。分类器可以是非常简单的，例如多数类投票分类器（Valley [59]）或一般的分类者，例如天真的贝叶斯（Nbiterative [60]）。 - *Binning:*此类别是指缺乏评估度量。这是通过创建指定数量的垃圾箱来离散属性的最简单方法。每个垃圾箱都被定义为先验，并分配每个属性的指定数量值。广泛使用的binning方法是相等的宽度和相等的频率。

3.1.2 Other Properties

我们可以评论与离散化有关的其他属性。它们还影响了通过离散器获得的操作和结果，但其程度低于上面所述的特征。此外，其中一些提出了各种各样的分类，可能会损害分类法的解释性。

- *Parametric vs. NonParametric:* 此属性是指通过离散器自动确定每个属性的间数。非验证者批准器计算考虑权衡的每个属性的适当数量的间隔数量

  在信息丢失或一致性之间以及获得最低数量之间。参数离散器需要用户固定的最大间隔数量。非参数离散器的示例是MDLP [16]和CAIM

  [57]。参数的示例是Chimerge [17]和CADD [52]。

- *Top-Down vs. Bottom Up:* 此属性仅以渐进的离散器为准。自上而下的方法以空离散为开头。它的改进过程仅仅是为离散化增加一个新的切口。另一方面，自下而上的方法以离散化开头，其中包含所有可能的切口。它的改进过程包括迭代合并两个间隔，从而删除了一个切口。经典的自上而下方法是MDLP [16]，一种众所周知的自下而上方法是Chimerge [17]。

- *Stopping Condition:* 这与用于停止离散过程的机制有关，必须在非参数方法中指定。众所周知的停止标准是最小描述长度度量[16]，置信阈值[17]或不一致比率[24]。

- *Disjoint vs. Non-Disjoint:* 分离方法将属性的值范围离散为拆分的内部服务，而无需重叠，而非分散方法将值范围限制为可能重叠的间隔。在此过程中审查的方法是不相交的，而模糊离散通常是非偏离的[36]。

- *Ordinal vs. Nominal:* 序数离散化trans-形式的定量数据引入序数定性数据，而名义离散化将其转化为名义定性数据，从而丢弃了有关顺序的信息。序数离散器不太常见，通常不被视为经典的离散剂[113]。

3.1.3 Criteria to Compare Discretization Methods

在比较离散方法时，有许多标准可用于评估每种算法的相对优势。这些包括间隔的数量，不一致，预测性分类率和时间要求

- *Number of Intervals:* 实用离散化的理想特征是离散的属性具有尽可能少的值，因为大量的热门服务可能会使学习缓慢且无效。 [28]。

- *Inconsistency:* 一种基于监督的度量，用于计算数据集中引起的不可避免的错误数量。一个不可避免的错误是与两个示例相关联的一个示例，该示例具有相同的输入属性和不同类标签的值。通常，具有连续属性的数据集是一致的，但是当在数据上应用离散化方案时，可能会获得不一致的数据集。这

表1：离散器

Complete name	Abbr. name	Reference	Complete name	Abbr. name	Reference
Equal Width Discretizer	**EqualWidth**	[61]	Self Organizing Map Discretizer	SOM-Disc	[62]
Equal Frequency Discretizer	**EqualFrequency**	[61]	Optimal Class-Dependent Discretizer	OCDD	[26]
No name specified	Chou91	[63]	*No name specified*	Butterworth04	[64]
Adaptive Quantizer	AQ	[65]	*No name specified*	Zhang04	[22]
Discretizer 2	D2	[28]	**Khiops**	**Khiops**	[66]
ChiMerge	**ChiMerge**	[17]	**Class-Attribute Interdependence Maximization**	**CAIM**	[57]
One-Rule Discretizer	**1R**	[27]	**Extended Chi2**	**Extended Chi2**	[67]
Iterative Dichotomizer 3 Discretizer	**ID3**	[10]	**Heterogeneity Discretizer**	**Heter-Disc**	[68]
Minimum Description Length Principle	**MDLP**	[16]	**Unsupervised Correlation Preserving Discretizer**	**UCPD**	[44]
Valley	Valley	[59], [69]	*No name specified*	Multi-MDL	[47]
Class-Attribute Dependent Discretizer	**CADD**	[52]	Difference Similitude Set Theory Discretizer	DSST	[70]
ReliefF Discretizer	ReliefF	[71]	Multivariate Interdependent Discretizer	MIDCA	[72]
Class-driven Statistical Discretizer	StatDisc	[14]	**MODL**	**MODL**	[20]
No name specified	NBIterative	[60]	Information Theoretic Fuzzy Partitioning	ITFP	[43]
Boolean Reasoning Discretizer	BRDisc	[21]	*No name specified*	Wu06	[73]
Minimum Description Length Discretizer	MDL-Disc	[74]	Fast Independent Component Analysis	FastICA	[75]
Bayesian Discretizer	**Bayesian**	[19]	Linear Program Relaxation	LP-Relaxation	[76]
No name specified	Friedman96	[77]	**Hellinger-Based Discretizer**	**HellingerBD**	[50]
Cluster Analysis Discretizer	**ClusterAnalysis**	[24]	**Distribution Index-Based Discretizer**	**DIBD**	[78]
Zeta	**Zeta**	[56]	Wrapper Estimation of Distribution Algorithm	WEDA	[53]
Distance-based Discretizer	**Distance**	[79]	Clustering + Rought Sets Discretizer	Cluster-RS-Disc	[25]
Finite Mixture Model Discretizer	FMM	[80]	**Interval Distance Discretizer**	**IDD**	[51]
Chi2	**Chi2**	[18]	**Class-Attribute Contingency Coefficient**	**CACC**	[58]
No name specified	FischerExt	[81]	Rectified Chi2	Rectified Chi2	[82]
Contextual Merit Numerical Feature Discretizer	CM-NFD	[83]	**Ameva**	**Ameva**	[84]
Concurrent Merger	ConMerge	[85]	Unification	Unification	[55]
Knowledge EXplorer Discretizer	KEX-Disc	[86]	Multiple Scanning Discretizer	MultipleScan	[87]
LVQ-based Discretization	LVQ-Disc	[88]	Optimal Flexible Frequency Discretizer	OFFD	[89]
No name specified	Multi-Bayesian	[90]	**Proportional Discretizer**	**PKID**	[12]
No name specified	A*	[91]	**Fixed Frequency Discretizer**	**FFD**	[12]
FUSINTER	**FUSINTER**	[54]	Discretization Class intervals Reduce	DCR	[92]
Cluster-based Discretizer	Cluster-Disc	[93]	MVD-CG	MVD-CG	[94]
Entropy-based Discretization According to Distribution of Boundary points	EDA-DB	[95]	Approximate Equal Frequency Discretizer	AEFD	[96]
No name specified	Clarke00	[97]	*No name specified*	Jiang09	[96]
Relative Unsupervised Discretizer	RUDE	[98]	Random Forest Discretizer	RFDisc	[99]
Multivariate Discretization	**MVD**	[45]	Supervised Multivariate Discretizer	SMD	[100]
Modified Learning from Examples Module	MODLEM	[101]	Clustering Based Discretization	CBD	[46]
Modified Chi2	**Modified Chi2**	[102]	Improved MDLP	Improved MDLP	[103]
HyperCluster Finder	HCF	[104]	Imfor-Disc	Imfor-Disc	[105]
Entropy-based Discretization with Inconsistency Checking	EDIC	[49]	Clustering ME-MDL	Cluster ME-MDL	[106]
Unparametrized Supervised Discretizer	**USD**	[107]	Effective Bottom-up Discretizer	EBDA	[108]
Rough Set Discretizer	RS-Disc	[109]	Contextual Discretizer	Contextual-Disc	[110]
Rough Set Genetic Algorithm Discretizer	RS-GA-Disc	[111]	**Hypercube Division Discretizer**	**HDD**	[48]
Genetic Algorithm Discretizer	GA-Disc	[112]			

离散器应获得的期望不一致水平为0.0。

- *Predictive Classification Rate:* 成功的算法通常能够离散训练设置，并显着降低了准备好处理数值数据的测试数据中学习者的预测能力。

- *Time requirements:* 仅在培训集中进行一次静态离散过程，因此这似乎不是非常重要的评估方法。但是，如果离散阶段花费的时间太长，对于实际应用来说可能会变得不切实际。在动态离散化中，按照学习者的要求重复多次操作，因此应有效地进行操作。

3.2离散方法和分类法

在写作时，文献中已经提出了80多种离散化方法。本节致力于根据本文遵循的标准列举和指定它们。在实验研究中，我们使用了30个离散因素，我们已经将其确定为最相关的那些。有关其描述的更多详细信息，读者可以访问与龙骨项目[1]相关的URL。其他，可以在龙骨软件[114]，[115]中找到这些算法的实现。

表1列出了本文回顾的离散器的列举。每个人都提供完整的名称，缩写和参考。本文确实如此

1. http://www.keel.es

由于空间限制，未收集离散器的描述。取而代之的是，我们建议读者咨询原始参考，以了解感兴趣的离散器的完整操作。实验研究中使用的离散器以粗体描述。该研究中使用的ID3离散器是C4.5中嵌入的众所周知的离散器的静态版本。

上面研究的属性可用于促进文献中提出的离散剂。所研究的七个特征使我们能够按照既定的顺序提出离散方法的分类法。 Table 1中列举的所有技术均在图2中绘制的分类法中收集。它说明了基于此顺序的层次结构之后的分类：静态/动态，Uni-variate/uni-variate/uni-variate/uni-variate/Multivariate，受监督/无人驾驶/无需进行的，分裂/合并/合并/合并/hybrid/hybrid，global/hybrid，global/hybrid，global/direct/direct/coremental和评估措施。选择该命令的基本原理是要明确表示分类法。

拟议的分类学为我们提供了许多离散化方法的组织，因此我们可以将其分类为类别并分析其行为。此外，我们可以强调分类法可以有用的其他方面。例如，它提供了现有方法和关系或相似之处的快照。它还描绘了家庭的大小，每个家庭所做的工作以及当前缺少的工作。最后，它为始于此主题或需要在实际应用程序中离散数据的研究人员/从业人员提供了有关离散化最先进的一般概述。

4实验框架，实证研究和结果分析

本节介绍了本文的实验框架，以及收集的结果以及对它们的讨论。第4.1小节将描述完整的实验设置。然后，我们对第4.2小节中使用的数据集获得的结果进行研究和分析。

4.1实验设置

本节的目的是显示与实验研究相关的所有属性和问题。我们指定数据集，验证过程，所使用的分类器，分类器和离散器的参数以及性能指标。用于对比结果的统计测试在本节末尾还评论了结果。

通过使用从UCI计算机学习数据库存储库[116]和龙骨数据集存储库[115] [2]获取的40个数据集，可以通过40个数据集进行离散化算法的性能。这些数据集的主要特征总结在表2中。对于每个数据集，名称，示例数，属性数量（NUMERIC和名义）和类数量。

2. http://www.keel.es/datasets.php

表2：分类数据集的摘要说明

Data Set	#Ex.	#Atts.	#Num.	#Nom.	#Cl.
abalone	4,174	8	7	1	28
appendicitis	106	7	7	0	2
australian	690	14	8	6	2
autos	205	25	15	10	6
balance	625	4	4	0	3
banana	5,300	2	2	0	2
bands	539	19	19	0	2
bupa	345	6	6	0	2
cleveland	303	13	13	0	5
contraceptive	1,473	9	9	0	3
crx	690	15	6	9	2
dermatology	366	34	34	0	6
ecoli	336	7	7	0	8
flare-solar	1066	9	9	0	2
glass	214	9	9	0	7
haberman	306	3	3	0	2
hayes	160	4	4	0	3
heart	270	13	13	0	2
hepatitis	155	19	19	0	2
iris	150	4	4	0	3
mammographic	961	5	5	0	2
movement	360	90	90	0	15
newthyroid	215	5	5	0	3
pageblocks	5,472	10	10	0	5
penbased	10,992	16	16	0	10
phoneme	5,404	5	5	0	2
pima	768	8	8	0	2
saheart	462	9	8	1	2
satimage	6,435	36	36	0	7
segment	2,310	19	19	0	7
sonar	208	60	60	0	2
spambase	4,597	57	57	0	2
specfheart	267	44	44	0	2
tae	151	5	5	0	3
titanic	2,201	3	3	0	2
vehicle	846	18	18	0	4
vowel	990	13	13	0	11
wine	178	13	13	0	3
wisconsin	699	9	9	0	2
yeast	1484	8	8	0	10

在这项研究中，已经使用了六个分类器，以发现离散器之间的性能差异。分类器是：

- *C4.5* [10]：众所周知的决策树，被认为是十大DM算法之一[9]。
- *DataSqueezer* [117]：该学习者属于归纳规则提取的家族。尽管具有相对简单性，但Datasqueezer还是一个非常有效的学习者。该算法产生的规则是紧凑且可理解的，但准确性在某种程度上是为了实现这一目标而降低的。
- *KNN*：基于一组对象之间相似性的最简单，最有效的方法之一。它也被认为是前10个DM算法之一[9]，它可以使用适当的距离功能（例如HVDM [118]）处理名义属性。它属于懒惰的学习家族[119]，[120]。
- *Naive Bayes*：这是前10个DM al-的另一个

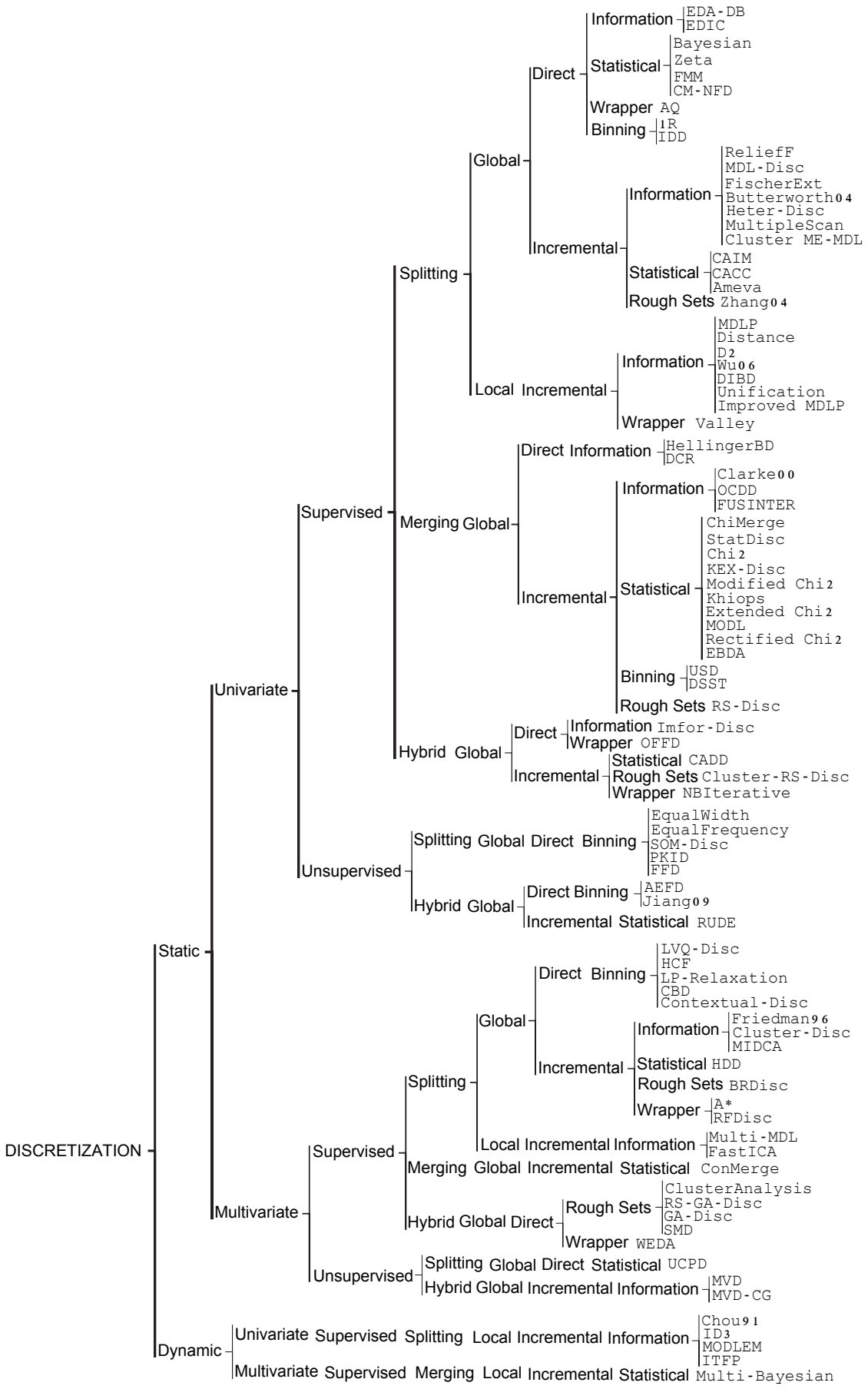


Fig. 2: Discretization Taxonomy

表3：离散器和分类器的参数

Method	Parameters
C4.5	pruned tree, confidence = 0.25, 2 examples per leaf
DataSqueezer	pruning and generalization threshold = 0.05
KNN	K=3, HVDM distance
PUBLIC	25 nodes between prune
Ripper	k=2, grow set = 0.66
1R	6 examples of the same class per interval
CADD	confidence threshold = 0.01
Chi2	inconsistency threshold = 0.02
ChiMerge	confidence threshold = 0.05
FDD	frequency size = 30
FUSINTER	$\alpha = 0.975$, $\lambda = 1$
HDD	coefficient = 0.8
IDD	neighborhood = 3, windows size = 3, nominal distance
MODL	optimized process type
UCPD	intervals = [3, 6], KNN map type, neighborhood = 6, minimum support = 25, merged threshold = 0.5, scaling factor = 0.5, use discrete

Gorithm [9]。它的目的是构建一个规则，该规则将使我们能够在建立概率时的属性独立性，将未来的对象分配给一个类。

- *PUBLIC* [121]：这是一个高级决策树，将修剪阶段与树的建筑阶段集成在一起，以避免扩大后来将修剪的分支。
- *Ripper* [122]：这是一种基于*separate and conquer*策略的广泛使用的规则感应方法。它结合了各种机制，以避免过度调整并同时处理数字和名义属性。获得的模型是决策列表的形式。

使用十倍跨验证（10-FCV）程序对所考虑的数据集进行分区。离散器和分类器的参数是其各自的作者推荐的参数。对于需要它们的那些方法，它们在表3中被指定。我们假设参数值的选择是由其自己的作者最佳选择的。然而，在要求将间隔数作为参数输入的离散器中，我们使用的是一个依赖数据集中实例数量的经验法则。它包括将实例数除以100，并在此结果和类数量之间取出最大值。所有的离散器和分类器都在每个分区中运行一次，因为它们是非策略的。

当多级分类问题解决时，两种绩效指标被广泛使用，因为它们的简单性和成功的应用。我们指的是准确性和Cohen的Kappa [123]措施，该措施将通过概括分类率来衡量效率离散器。

- *Accuracy*：相对于分类总数，成功的命中次数是否相对于分类数量。多年来，它一直是评估分类器性能的最常用度量[2]，[124]。
- *Cohen's kappa*：是*accuracy*的替代方法，一种方法，

表4：离散剂的内在证券收集的平均结果：获得的间隔数量和训练和测试数据中的不一致率

Number Int.		*Incons. Train*		*Incons. Tst*	
Heter-Disc	8.3125	ID3	0.0504	ID3	0.0349
MVD	18.4575	PKID	0.0581	PKID	0.0358
Distance	23.2125	Modified Chi2	0.0693	FFD	0.0377
UCPD	35.0225	FFD	0.0693	HDD	0.0405
MDLP	36.6600	HDD	0.0755	Modified Chi2	0.0409
Chi2	46.6350	USD	0.0874	USD	0.0512
FUSINTER	59.9850	ClusterAnalysis	0.0958	Khiops	0.0599
DIBD	64.4025	Khiops	0.1157	ClusterAnalysis	0.0623
CADD	67.7100	EqualWidth	0.1222	EqualWidth	0.0627
ChiMerge	69.5625	EqualFrequency	0.1355	EqualFrequency	0.0652
CAIM	72.5125	Chi2	0.1360	Chi2	0.0653
Zeta	75.9325	Bayesian	0.1642	FUSINTER	0.0854
Ameva	78.8425	MODL	0.1716	MODL	0.0970
Khiops	130.3000	FUSINTER	0.1735	HellingerBD	0.1054
1R	162.1925	HellingerBD	0.1975	Bayesian	0.1139
EqualWidth	171.7200	IDD	0.2061	UCPD	0.1383
Extended Chi2	205.2650	ChiMerge	0.2504	ChiMerge	0.1432
HellingerBD	244.6925	UCPD	0.2605	IDD	0.1570
EqualFrequency	267.7250	CAIM	0.2810	CAIM	0.1589
PKID	295.9550	Extended Chi2	0.3048	Extended Chi2	0.1762
MODL	335.8700	Ameva	0.3050	Ameva	0.1932
FFD	342.6050	1R	0.3112	CACC	0.2047
IDD	349.1250	CACC	0.3118	1R	0.2441
Modified Chi2	353.6000	MDLP	0.3783	Zeta	0.2454
CACC	505.5775	Zeta	0.3913	MDLP	0.2501
ClusterAnalysis	1116.1800	MVD	0.4237	DIBD	0.2757
USD	1276.1775	Distance	0.4274	Distance	0.2987
Bayesian	1336.0175	DIBD	0.4367	MVD	0.3171
ID3	1858.3000	CADD	0.6532	CADD	0.5688
HDD	2202.5275	Heter-Disc	0.6749	Heter-Disc	0.5708

数十年来闻名，它可以补偿随机命中[123]。其最初的目的是衡量两个人观察相同现象的人之间的共识或分歧程度。 Cohen的Kappa可以适应分类任务，因此建议使用其使用，因为它以与AUC度量相同的方式将随机的成功作为标准视为[125]。

计算Cohen Kappa的一种简单方法是在分类任务中利用所产生的混乱矩阵。特别是，可以使用以下表达来获得科恩的kappa措施：

$$kappa = \frac{N\sum_{i=1}^{C} y_{ii} - \sum_{i=1}^{C} y_{i.}y_{.i}}{N^2 - \sum_{i=1}^{C} y_{i.}y_{.i}},$$

其中y_{ii}是所得混淆矩阵的主要对角线中的细胞计数，N是示例的数量，C是类值的数量，$y_{.i}$，$y_{.i}$，$y_{i.}$，$y_{i.}$分别是列的混淆矩阵的总数。科恩的kappa范围从– 1（总分歧）到0（ran-dom分类）到1（完美的协议）。作为标量，将其应用于二进制二级时，它比ROC曲线的表现力较低。但是，对于多类问题，Kappa是一个非常有用但简单的仪表，用于测量分类器的准确性，同时补偿随机成功。

实证研究涉及30个离散化甲基

表1中列出的ODS。我们想概述实现仅基于各个作者在论文中给出的描述和规范。

统计分析将通过非参数统计检验进行。在[126]，[127]，[128]中，Authors建议一组简单，安全且可靠的非参数测试，用于分类器的统计比较。 Wilcoxon测试[129]将用于研究研究中考虑的所有离散剂之间的成对比较。有关这些统计过程的更多信息，可以在
Statistical Inference in Computational Intelligence and Data Mining [3]上的Sci2S主题公共网站上找到用于机器学习领域的专门设计的信息。

4.2分析和经验结果

表4列出了与40个数据集中所有离散器的培训和测试数据中间隔数量和不一致率相对应的平均结果。类似地，表5和6收集了每个考虑的分类人员的准确性和KAPPA措施的平均结果。对于每个度量，将离散器从最好的到最差。在表5和6中，我们突出显示了那些在每种度量中最佳和最差方法之间的性能范围内的范围的5%以内，即$value_{best}-(0.05\cdot(value_{best}-value_{worst}))$。无论其在表格中的特定位置如何，它们都应被视为每个类别中的出色方法。

每个数据集，离散器和分类器（包括平均和标准偏差）的所有详细结果都可以在URL http://sci2s.ugr.es/ iveptization上找到。为了紧凑，我们将在本文中包括并分析总结结果。

在本研究中采用了Wilcoxon测试[129]，[126]，[127]，考虑到等于$\alpha = 0.05$的意义水平。表7、8和9显示了在所有离散器和措施中Wilcoxon检验中涉及的所有可比性比较的摘要，分别是间隔和不一致率，准确性和KAPPA的数量。同样，上述URL中提到的所有可能的离散器之间的单个比较，在上面提到的URL中，可以找到每种度量和分类器的统计结果的详细报告。本文中的表（7、8和9）总结了行中的每种方法，使用'+'符号表示的列下的Wilcoxon测试的分散数量优于使用Wilcoxon测试。带有'±'符号的列指示该方法中该方法获得的胜利和联系数。每列的最大值由阴影单元突出显示。

最后，为了说明平均结果差异的幅度以及每个离散器所产生的间隔数量与每个分类器获得的精度之间的关系，图3描述了A

3. http://sci2s.ugr.es/sicidm/

表7：Wilcoxon测试会导致间隔和不一致的数量

	N. Intervals		Incons. Tra		Incons. Tst	
	+	±	+	±	+	±
1R	10	21	3	17	2	20
Ameva	13	21	6	16	4	21
Bayesian	2	4	10	29	7	29
CACC	7	22	4	17	4	21
CADD	21	28	0	1	0	1
CAIM	14	23	6	19	6	20
Chi2	15	26	9	20	9	20
ChiMerge	15	23	6	20	6	23
ClusterAnalysis	1	4	15	29	9	29
DIBD	21	27	2	7	2	8
Distance	26	28	2	6	2	6
EqualFrequency	7	12	12	26	11	29
EqualWidth	11	18	16	26	13	29
Extended Chi2	14	27	2	14	2	18
FFD	5	8	21	29	16	29
FUSINTER	14	22	11	23	8	29
HDD	0	2	18	29	14	29
HellingerBD	9	15	8	21	7	26
Heter-Disc	29	29	0	1	0	1
ID3	0	1	23	29	16	29
IDD	5	11	8	28	6	29
Khiops	9	15	15	27	12	29
MDLP	22	27	3	9	3	11
Modified Chi2	7	13	17	26	15	29
MODL	5	14	12	24	7	29
MVD	23	28	2	13	2	13
PKID	5	8	22	29	16	29
UCPD	17	25	6	17	5	20
USD	2	4	18	29	15	29
Zeta	12	23	3	9	3	13

每个分类器的X-Y轴图形所反映的平均间隔数量和准确性之间的对抗。它还可以帮助我们查看离散化行为的差异，当时它被用于不同的分类器。

一旦在提到的表和图形中提出结果后，我们就可以强调从中观察到的一些有趣的属性，我们可以指出表现最好的离散器：

- 关于间隔的数量，将数值属性划分为较少的离散器是*Heter-Disc*，*MVD*和*Distance*，而需要大量切割点的离散化值为*HDD*，*ID3*，*ID3*和*Bayesian*。 Wilcoxon测试证实了*Heter-Disc*是获得最小间隔优于其余间隔的离散器。

- 培训数据和测试数据中的不一致率遵循所有离散器的类似趋势，

表5：考虑到六个分类器的准确度的平均结果

C4.5		*DataSqueezer*		*KNN*		*Naive Bayes*		*PUBLIC*		*Ripper*	
FUSINTER	0.7588	Distance	0.5666	PKID	0.7699	PKID	0.7587	FUSINTER	0.7448	Modified Chi2	0.7241
ChiMerge	0.7494	CAIM	0.5547	FFD	0.7594	Modified Chi2	0.7578	CAIM	0.7420	Chi2	0.7196
Zeta	0.7488	Ameva	0.5518	Modified Chi2	0.7573	FUSINTER	0.7576	ChiMerge	0.7390	PKID	0.7097
CAIM	0.7484	MDLP	0.5475	EqualFrequency	0.7557	ChiMerge	0.7543	MDLP	0.7334	MODL	0.7089
UCPD	0.7447	Zeta	0.5475	Khiops	0.7512	FFD	0.7535	Distance	0.7305	FUSINTER	0.7078
Distance	0.7446	ChiMerge	0.5472	EqualWidth	0.7472	CAIM	0.7535	Zeta	0.7301	Khiops	0.6999
MDLP	0.7444	CACC	0.5430	FUSINTER	0.7440	EqualWidth	0.7517	Chi2	0.7278	FFD	0.6970
Chi2	0.7442	Heter-Disc	0.5374	ChiMerge	0.7389	Zeta	0.7507	UCPD	0.7254	EqualWidth	0.6899
Modified Chi2	0.7396	DIBD	0.5322	CAIM	0.7381	EqualFrequency	0.7491	Modified Chi2	0.7250	EqualFrequency	0.6890
Ameva	0.7351	UCPD	0.5172	MODL	0.7372	MODL	0.7479	Khiops	0.7200	CAIM	0.6870
Khiops	0.7312	MVD	0.5147	HellingerBD	0.7327	Chi2	0.7476	Ameva	0.7168	HellingerBD	0.6816
MODL	0.7310	FUSINTER	0.5126	Chi2	0.7267	Khiops	0.7455	HellingerBD	0.7119	USD	0.6807
EqualFrequency	0.7304	Bayesian	0.4915	USD	0.7228	USD	0.7428	EqualFrequency	0.7110	ChiMerge	0.6804
EqualWidth	0.7252	Extended Chi2	0.4913	Ameva	0.7220	ID3	0.7381	MODL	0.7103	ID3	0.6787
HellingerBD	0.7240	Chi2	0.4874	ID3	0.7172	Ameva	0.7375	CACC	0.7069	Zeta	0.6786
CACC	0.7203	HellingerBD	0.4868	ClusterAnalysis	0.7132	Distance	0.7372	DIBD	0.7002	HDD	0.6700
Extended Chi2	0.7172	MODL	0.4812	Zeta	0.7126	MDLP	0.7369	EqualWidth	0.6998	Ameva	0.6665
DIBD	0.7141	CADD	0.4780	HDD	0.7104	ClusterAnalysis	0.7363	Extended Chi2	0.6974	UCPD	0.6651
FFD	0.7091	EqualFrequency	0.4711	UCPD	0.7090	HellingerBD	0.7363	HDD	0.6789	CACC	0.6562
PKID	0.7079	1R	0.4702	MDLP	0.7002	HDD	0.7360	FFD	0.6770	Extended Chi2	0.6545
HDD	0.6941	EqualWidth	0.4680	Distance	0.6888	UCPD	0.7227	PKID	0.6758	Bayesian	0.6521
USD	0.6835	IDD	0.4679	IDD	0.6860	Extended Chi2	0.7180	USD	0.6698	ClusterAnalysis	0.6464
ClusterAnalysis	0.6813	USD	0.4651	Bayesian	0.6844	CACC	0.7176	Bayesian	0.6551	MDLP	0.6439
ID3	0.6720	Khiops	0.4567	CACC	0.6813	Bayesian	0.7167	ClusterAnalysis	0.6477	Distance	0.6402
1R	0.6695	Modified Chi2	0.4526	DIBD	0.6731	DIBD	0.7036	ID3	0.6406	IDD	0.6219
Bayesian	0.6675	HDD	0.4308	1R	0.6721	IDD	0.6966	MVD	0.6401	Heter-Disc	0.6084
IDD	0.6606	ClusterAnalysis	0.4282	Extended Chi2	0.6695	1R	0.6774	IDD	0.6352	1R	0.6058
MVD	0.6499	PKID	0.3942	MVD	0.6062	MVD	0.6501	1R	0.6332	DIBD	0.5953
Heter-Disc	0.6443	ID3	0.3896	Heter-Disc	0.5524	Heter-Disc	0.6307	Heter-Disc	0.6317	MVD	0.5921
CADD	0.5689	FFD	0.3848	CADD	0.5064	CADD	0.5669	CADD	0.5584	CADD	0.4130

表6：Kappa的平均结果考虑了六个分类器

C4.5		*DataSqueezer*		*KNN*		*Naive Bayes*		*PUBLIC*		*Ripper*	
FUSINTER	0.5550	CACC	0.2719	PKID	0.5784	PKID	0.5762	CAIM	0.5279	Modified Chi2	0.5180
ChiMerge	0.5433	Ameva	0.2712	FFD	0.5617	Modified Chi2	0.5742	FUSINTER	0.5204	Chi2	0.5163
CAIM	0.5427	CAIM	0.2618	Modified Chi2	0.5492	FUSINTER	0.5737	ChiMerge	0.5158	MODL	0.5123
Zeta	0.5379	ChiMerge	0.2501	Khiops	0.5457	FFD	0.5710	MDLP	0.5118	FUSINTER	0.5073
MDLP	0.5305	FUSINTER	0.2421	EqualFrequency	0.5438	ChiMerge	0.5650	Distance	0.5074	Khiops	0.4939
UCPD	0.5299	UCPD	0.2324	EqualWidth	0.5338	Chi2	0.5620	Zeta	0.5010	PKID	0.4915
Ameva	0.5297	Zeta	0.2189	CAIM	0.5260	CAIM	0.5616	Ameva	0.4986	EqualFrequency	0.4892
Chi2	0.5290	USD	0.2174	FUSINTER	0.5242	EqualWidth	0.5593	Chi2	0.4899	ChiMerge	0.4878
Distance	0.5288	Distance	0.2099	ChiMerge	0.5232	Khiops	0.5570	UCPD	0.4888	EqualWidth	0.4875
Modified Chi2	0.5163	Khiops	0.2038	MODL	0.5205	EqualFrequency	0.5564	Khiops	0.4846	CAIM	0.4870
MODL	0.5131	HDD	0.2030	HellingerBD	0.5111	MODL	0.5564	CACC	0.4746	Ameva	0.4810
EqualFrequency	0.5108	EqualFrequency	0.2016	Chi2	0.5100	USD	0.5458	HellingerBD	0.4736	FFD	0.4809
Khiops	0.5078	HellingerBD	0.1965	Ameva	0.5041	Zeta	0.5457	Modified Chi2	0.4697	Zeta	0.4769
HellingerBD	0.4984	Bayesian	0.1941	USD	0.4943	Ameva	0.5456	MODL	0.4620	HellingerBD	0.4729
CACC	0.4961	MODL	0.1918	HDD	0.4878	ID3	0.5403	EqualFrequency	0.4535	USD	0.4560
EqualWidth	0.4909	MDLP	0.1875	ClusterAnalysis	0.4863	HDD	0.5394	DIBD	0.4431	UCPD	0.4552
Extended Chi2	0.4766	PKID	0.1846	Zeta	0.4831	MDLP	0.5389	EqualWidth	0.4386	CACC	0.4504
DIBD	0.4759	ID3	0.1818	ID3	0.4769	Distance	0.5368	Extended Chi2	0.4358	MDLP	0.4449
FFD	0.4605	EqualWidth	0.1801	UCPD	0.4763	HellingerBD	0.5353	HDD	0.4048	Distance	0.4429
PKID	0.4526	Modified Chi2	0.1788	MDLP	0.4656	ClusterAnalysis	0.5252	FFD	0.3969	HDD	0.4403
HDD	0.4287	DIBD	0.1778	Distance	0.4470	UCPD	0.5194	PKID	0.3883	ID3	0.4359
USD	0.4282	Chi2	0.1743	CACC	0.4367	CACC	0.5128	USD	0.3845	Extended Chi2	0.4290
ClusterAnalysis	0.4044	IDD	0.1648	IDD	0.4329	Extended Chi2	0.4910	MVD	0.3461	ClusterAnalysis	0.4252
ID3	0.3803	FFD	0.1635	Extended Chi2	0.4226	Bayesian	0.4757	ClusterAnalysis	0.3453	Bayesian	0.3987
IDD	0.3803	ClusterAnalysis	0.1613	Bayesian	0.4201	DIBD	0.4731	Bayesian	0.3419	DIBD	0.3759
MVD	0.3759	Extended Chi2	0.1465	DIBD	0.4167	IDD	0.4618	ID3	0.3241	IDD	0.3650
Bayesian	0.3716	MVD	0.1312	1R	0.3940	1R	0.3980	IDD	0.3066	MVD	0.3446
1R	0.3574	1R	0.1147	MVD	0.3429	MVD	0.3977	1R	0.3004	1R	0.3371
Heter-Disc	0.2709	Heter-Disc	0.1024	Heter-Disc	0.2172	Heter-Disc	0.2583	Heter-Disc	0.2570	Heter-Disc	0.2402
CADD	0.1524	CADD	0.0260	CADD	0.1669	CADD	0.1729	CADD	0.1489	CADD	0.1602

考虑到测试数据中获得的不一致始终低于培训数据中。*ID3*是培训和测试数据中获得最低平均不一致率的离散器，尽管

Wilcoxon测试无法找到IT和其他两个离散器之间的明确差异：*FFD*和*PKID*。我们可以观察到的间隔数量之间的密切关系，

表8：Wilcoxon测试可导致准确性

	C4.5		Data Squeezer		KNN		Naive Bayes		PUBLIC		Ripper	
	+	±	+	±	+	±	+	±	+	±	+	±
1R	1	12	3	23	2	19	1	9	1	12	1	11
Ameva	14	29	17	29	8	26	9	29	13	29	9	29
Bayesian	1	9	5	26	2	12	2	11	0	11	2	17
CACC	9	28	16	29	2	18	5	28	9	29	4	26
CADD	0	1	1	22	0	1	0	1	0	6	0	0
CAIM	16	29	16	29	11	28	10	29	16	29	11	28
Chi2	13	29	4	26	6	27	9	29	11	29	19	29
ChiMerge	17	29	18	29	13	28	10	29	17	29	9	28
ClusterAnalysis	1	10	0	12	5	24	6	27	1	11	2	20
DIBD	6	21	8	29	2	9	2	8	9	23	1	5
Distance	13	29	16	29	2	17	7	26	13	28	2	13
EqualFrequency	10	27	3	21	18	29	9	29	10	26	11	27
EqualWidth	7	20	2	18	11	28	8	29	6	20	9	27
Extended Chi2	9	27	4	26	3	19	3	17	6	25	2	25
FFD	5	15	0	5	20	28	8	29	1	13	10	27
FUSINTER	21	29	9	29	12	28	15	29	20	29	11	29
HDD	1	18	0	14	4	23	5	28	0	24	7	26
HellingerBD	10	27	4	22	7	26	7	28	10	26	6	26
Heter-Disc	0	9	9	29	0	2	0	3	0	11	1	10
ID3	1	10	0	5	5	22	4	28	0	11	5	26
IDD	1	10	3	23	4	21	2	14	0	12	1	16
Khiops	12	27	3	18	18	29	9	29	9	27	11	29
MDLP	14	29	14	29	3	22	8	29	15	29	2	16
Modified Chi2	11	27	3	21	17	28	10	29	9	29	23	29
MODL	12	28	5	23	14	28	9	29	10	28	17	29
MVD	1	15	5	29	1	8	1	7	0	19	1	13
PKID	5	15	0	6	27	29	9	29	1	13	15	29
UCPD	14	29	7	26	4	17	2	15	14	28	3	19
USD	1	13	3	19	6	23	6	29	1	19	7	25
Zeta	14	29	17	29	4	20	9	29	14	29	7	27

表9：Wilcoxon测试结果Kappa

	C4.5		Data Squeezer		KNN		Naive Bayes		PUBLIC		Ripper	
	+	±	+	±	+	±	+	±	+	±	+	±
1R	1	11	0	15	2	16	2	8	1	13	1	11
Ameva	15	29	24	29	11	26	11	29	16	29	9	29
Bayesian	1	8	1	24	2	10	2	8	1	11	2	17
CACC	11	28	25	29	3	16	7	25	13	29	4	26
CADD	0	1	0	3	0	1	0	1	0	2	0	0
CAIM	17	29	22	29	13	28	11	29	21	29	11	28
Chi2	14	29	2	24	11	27	10	29	13	29	19	29
ChiMerge	19	29	22	29	13	28	11	29	18	29	9	28
ClusterAnalysis	2	10	2	21	5	23	6	22	1	11	2	20
DIBD	8	20	1	24	2	10	2	7	7	18	1	5
Distance	16	29	1	26	2	16	7	28	16	29	2	13
EqualFrequency	11	25	3	25	18	29	10	29	10	23	11	27
EqualWidth	7	20	2	23	14	27	8	28	6	18	9	27
Extended Chi2	10	27	1	20	2	17	3	16	6	23	2	25
FFD	6	14	1	19	23	28	12	29	2	14	10	27
FUSINTER	21	29	16	29	14	28	18	29	19	29	11	29
HDD	2	17	5	25	5	22	6	25	1	22	7	26
HellingerBD	11	23	4	23	9	26	7	21	11	24	6	26
Heter-Disc	0	6	0	12	0	2	0	2	0	8	1	10
ID3	1	8	2	22	4	20	6	26	0	10	5	26
IDD	1	9	1	23	2	18	2	15	1	11	1	16
Khiops	11	24	5	24	18	29	10	29	13	25	11	29
MDLP	16	29	1	24	6	22	8	29	19	29	2	16
Modified Chi2	12	27	1	21	17	27	14	29	9	28	23	29
MODL	12	28	4	24	14	27	12	29	11	28	17	29
MVD	1	12	0	19	1	10	1	6	1	16	1	13
PKID	5	14	2	23	27	29	14	29	2	14	15	29
UCPD	14	29	15	28	4	16	4	16	13	25	3	19
USD	4	13	9	25	6	23	6	25	3	15	7	25
Zeta	15	29	9	27	3	18	6	27	16	29	7	27

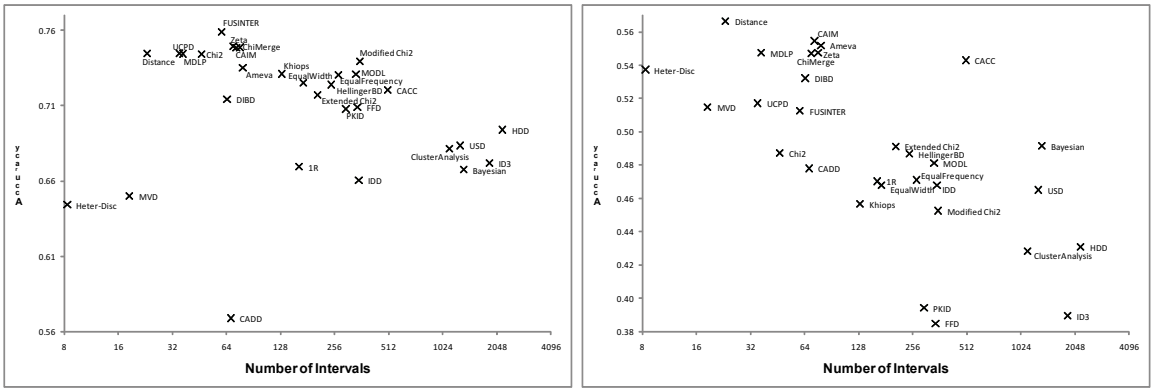
不一致的率，即计算较少切口点的离散率通常是较不一致的速率的不一致率。他们有可能具有数据一致性以简化结果，尽管一致性通常与准确性不相关，正如我们将在下面看到的那样。

- 在决策树（*C4.5*和*PUBLIC*）中，可以将离散剂视为表现最好的人。考虑到平均精度，*FUSINTER*，*ChiMerge*和*CAIM*在其余部分中脱颖而出。此子集还添加了平均kappa，*Zeta*和*MDLP*的平均值。 Wilcoxon测试确认了这一结果，并添加了另一个离散器*Distance*，该*Distance*的表现优于29种方法中的16个。强调的所有方法均受监督，增量（*Zeta*），并使用统计和信息度量作为评估者。分裂/合并和本地/全球属性对决策树没有影响。

- 考虑规则归纳（*DataSqueezer*和*Rip- per*），最佳性能的离散器是*Distance*，*Modified Chi2*，*Chi2*，*Chi2*，*PKID*}和*MODL*的平均准确性和*MODL*在准确性和*CACC*，v22}，v23}，v24中{卡帕。在这种情况下，由于Wilcoxon测试强调*ChiMerge*是*DataSqueezer*而不是*Distance*的最佳性能，并且在子集中合并*Zeta*，因此结果非常不规则。使用*Ripper*，Wilcoxon测试确认通过平均准确性和kappa获得的结果。很难辨别一组通用的属性集，这些属性定义了最佳性能的离散剂，这是因为规则诱导方法在其操作上的不同程度与决策树更大的程度不同。但是，我们可以指出的是，在最佳方法的子集中，统计评估中占主导地位和受监督的离散剂。

- 懒惰和贝叶斯学习可以进行分析，因为KNN中使用的HVDM距离与考虑到属性的差异的计算高度相关[118]。关于懒惰和贝叶斯学习，*KNN*和*Naive Bayes*，可重新分散器的子集由*PKID*，*FFD*，*Mod-*，*Modified Chi2*，*ified Chi2*，*FUSINTER*，*FUSINTER*，*ChiMerge*，*ChiMerge*，*ChiMerge*，{被使用；和*Chi2*，*Khiops*，*EqualFrequency*和*MODL*时必须在考虑普通kappa时添加。 Wilcoxon的统计报告向我们告知我们存在两种出色的方法：*PKID*的*PKID*，对于*KNN*，对于*Naive Bayes*而言，*KNN*的表现优于27/29和*FUSINTER*。在这里，受到监督和无监督，直接和内置，安装和统计/信息评估是最佳培养方法中存在的特征。但是，我们可以看到它们都是全球的，从而确定了构成分类方法的趋势。

- 通常，由离散剂进行的准确性和Kappa绩效没有太大差异。考虑到Kappa的差异通常由于其提供的随机成功赔偿，这两个评估指标中的差异都非常相似。令人惊讶的是，在*DataSqueezer*中，准确性和kappa提供了最大的行为差异，但它们的动机是因为该方法着重于获得简单的规则集，



(a) C4.5 (b) DataSqueezer

(c) KNN (d) Naive Bayes

(e) PUBLIC (f) Ripper

图3：准确性与间隔数

在后台留下精度。

- 显然，离散化与所使用的分类器之间存在直接依赖性。我们已经指出，可以检测到决策树和懒惰/贝叶斯学习的类似行为，而在规则归纳学习中，算法的运行条件条件是离散剂的有效性。了解每种类型的离散器的合适离散器的子集是理解和提出改进该地区的良好起点。

- 可以就准确性与数量之间的关系做出另一个有趣的评论

文已被PU接受

间隔者得出的间隔。图3支持以下假设：它们之间没有直接相关。计算几个切口点的离散器不必在确定性方面获得差的结果，反之亦然。图3A，3C，3D和3E指出，由*Distance*计算出的切割点给出的间隔数量有最低限制以保证准确的模型。图3b显示了*DataSqueezer*的时间增加，随着间隔数量的增加，这是分类器的固有行为。

- Fina Lly，我们可以强调全球最佳的一部分 本

考虑到获得的间隔数量和获得的准确性之间的权衡。在此子集中，我们可以包括*FUSINTER*，*Distance*，*Chi2*，*MDLP*和*UCPD*。

另一方面，以所研究的30个离散剂为中心的分析如下：

- 许多经典的离散剂通常是最好的分散剂。 *ChiMerge*，*MDLP*，*Zeta*，*Distance*和*Chi2*就是这种情况。
- 考虑到这些年来，其他经典的离散剂不如应有的好：*EqualWidth*，*EqualFrequency*，*1R*，*ID3* (，*ID3* (静态版本要比C4.5操作中的动态插入)，)，v18}，*CADD*}，v10}}和*Bayesian*和*Bayesian*和{
- 经典方法的轻微修改大大增强了它们的结果，例如*FUS- INTER*，*Modified Chi2*，*PKID*和*FFD*;但是在其他情况下，扩展使它们的孔子降低：*USD*，*Extended Chi2*。
- 在不利的情况下进行了评估的有希望的技术是*MVD*和*UCP*，这些方法是无监督的方法，可用于应用于其他DM问题，除了分类。
- 与经典方法相比，最近提出的方法是竞争性的，甚至在某些sce-narios中胜过它们，*Khiops*，*CAIM*，*MODL*，*MODL*，*Ameva*}和*CACC*。但是，通常报告结果不良结果的最新建议是*Heter-Disc*，*HellingerBD*，*DIBD*，*IDD*和*HDD*。
- 最后，这项研究涉及比以前工作中考虑的数量更高的数据集，并且得出的结论是对特定离散器的公正性。但是，我们必须强调与这些先前作品的结论一些巧合。例如，在[102]中，作者就准确性提出了改进的*Chi2*的改进版本，从而删除了用户参数选择。我们检查并测量实际的改进。在[12]中，作者对*Naive Bayes*进行了一项激烈的理论和态度研究，并根据他们的结论提出*PKID*和*FFD*。在本文中，我们证实*PKID*是*Naive Bayes*甚至*KNN*的最佳方法。最后，我们可能会注意到*CAIM*是最简单的疾病之一，在本研究中也显示了其有效性。

5总结和全球指南

本文对文献中提出的离散方法进行了详尽的调查。已经研究了Ba-sic和高级属性，现有的工作和相关领域。基于研究的主要特征，我们设计了一种分类方法的分类法。此外，最重要的

在大量的分类数据集上，已对离散器（经典和近期）进行了经验分析。为了加强研究，已经添加了基于非参数测试的统计分析，以支持得出的结论。可以提出几种评论和准则：

- 有兴趣应用离散化方法感兴趣的研究人员/从业人员应意识到定义它们的规定，以便在每种情况下选择最合适的方法。制定的分类学和实证研究可以帮助做出这一决定。
- 在提出新的离散剂的提议中，应在综合研究中使用最好的副业和与新提案的基本符合条件的提案。为了做到这一点，分类法和结果分析可以正确地指导未来的建议。
- 本文有助于非专业的离散化，以区分方法，做出有关其应用和理解其行为的决定。
- 重要的是要了解每种裁判率的主要优势。在本文中，许多离散因素已经经过经验分析，但是我们不能就表现最好的结论给出一个罪恶的结论。这取决于解决的问题和所使用的数据挖掘方法，但是此处提供的结果可能有助于限制候选人集。
- 实证研究使我们能够在整个集合中强调几种方法： - *FUSINTER*，*ChiMerge*，*CAIM*和*Modified Chi2*，考虑到所有类型的分类器，都提供出色的性能。 - *PKID*，*FFD*是懒惰和贝叶斯学习的合适方法，*CACC*，*Distance*和*MODL*是规则归纳学习中的不错选择。 - *FUSINTER*，*Distance*，*Chi2*，*MDLP*和*UCPD*在产生的间隔和准确性之间获得令人满意的权衡。

希望决定要确定哪种离散化计划的研究人员/从业人员需要知道本文或数据的实验将如何有益并指导他/她。随着未来的工作，我们提出了分类学中有关某些数据特征的每个属性的分析，例如标签数量，尺寸或原始属性的动态范围。遵循这种趋势，我们期望考虑到数据集的某些基本特征，将找到最合适的离散器。

致谢

这项工作得到了研究项目TIN2011-28488和TIC-6858的支持。 J.A. S ´Aez和V. Lopez拥有西班牙部的FPU奖学金

教育。作者要感谢评论者的建议。

参考

[1] J. Han，M。Kamber和J. Pei，*Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems)*。 Morgan Kaufmann，2006。[2] I. H. Witten，E。F rank和M. A. Hall，*Data Mining: Practical machine learning tools and techniques. 3rd Edition.*
machine learning tools and techniques. 3rd Edition. Morgan Kauf-Mann，2 011。[3] I. Kononenko和M. Kukar和M. Kukar，
Machine Learning and Data Mining:
Introduction to Principles and Algorithms。 Horwood Publishing Limited ，2007年。[4] K. J. Cios，W。Pedrycz，R。W。Swiniarski和L. A. Kurga n，*Data Mining: A Knowledge Discovery Approach*。 Springer，2007。[5] D. Pyle，*Data preparation for data mining*。 Morgan Kaufmann Publishers Inc.，1999。[6] H. Liu，F。Hussain，C。L. Tan和M. Dash，“离散化：一种启示技术”，*Data Mining and Knowledge Discovery*，第1卷。 6，不 。 4，pp。393–423，2002。[7] J. Dougherty，R。Kohavi和M. Sahami，“被监督和未监督连续特征的离散化”，载于*Proceedings of the Twelfth International Conference onMachine Learning (ICML)*，1995 年，第194-202页。 [8] Y. Yang，G。I. Webb和X. Wu，“离散方法”，在*Data Mining and Knowledge Discovery Handbook*中，2010年，第101-1 16页。 [9] X. Wu和V. Kumar编辑，
The Top Ten Algorithms in Data Mining。 Chapman＆Hall/CRC数据挖掘与知识发现，2009年。[10] J. R. Quinlan，
C4.5: programs for machine learning。 Morgan Kaufmann Publishers Inc. ，1993。[11] R. Agrawal和R. Srikant，“采矿协会规则的快速算法”，载于*Proceedings of the 20th Very Large Data Bases conference (VLDB)*，19 94年，第487-499页。 [12] Y. Yang和G. I. Webb，“幼稚巴约斯学习的离散化：管理离散化偏见和差异”，*Machine Learning*，第1卷。 74，不。 1，第39–74页，2009年。[13] M. J. Flores，J。A。G ´Amez，A。M。Ma rt ´nez和J. M. Puerta，“处理数字属性时，比较了贝叶斯净工程分类器时：bayesian net-Net-net-net-firfirers：converiation net-net-net-firfirers fielus c lassie classifirers：” *Applied Intelligence, in press DOI: 10.1007/s10489-011-0286-z*，2011年。[14] M. Richeldi和M. Rossotto，“连续属性的类统计统计分配”，in *Proceedings of the 8th European Conference on Machine Learning (ECML)* ，ser。 ECML ’95，1995，第335–338页。 [15] B. Chlebus和S. H. Nguye n，“在*Lecture Notes in Artificial Intelligence*中找到两个属性的最佳置换 ”，第1卷。 1424，1998，第537–544页。 [16] U. M. Fayyad和K. B. Irani ，“分类学习的连续值属性的多间隔离散化”，载于 *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI)*，1993，第1022-1029页。 [17] R. Kerber，“ Chimerg e：数字属性的离散化”，
National Conference on Artifical Intelligence American Association for Artificial Intelligence (AAAI)，1992年，第123–128页。 [18] H. Liu和R. S etiono，“通过离散化的特征选择”，
IEEE Transactions on Knowledge and Data Engineering，第1卷。 9，第6 42–645页，1997年。[19] 39，pp。688–691，1996。[20] M. boull ´e，“ modl：连续属性的贝叶斯最佳离散方法”，*Machine Learning*，第1卷。 65，不。 1，第131-165页，2006年。[21] S. H. Nguyen和A. Skowron，“真实价值属性的量化 - 粗糙集和布尔推理方法”，载于 *Proceedings of the Second Joint Annual Conference on Information Sciences (JCIS)*，1995年，第34-37页。 [22] G. Zhang，L。Hu和W. Jin，“在粗糙集理论及其应用中连续属性的离散化”，载于*Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems of the 2004 IEEE Conference on Cybernetics and Intelligent Systems (CIS)* ，2004年，第1020-1026页。 [23] A. A. Bakar，Z。A. Othman和N. L. M. Shuib，“为数据离散技术构建新的分类法”，*Proceedings on Conference on Data Mining and Optimization (DMO)*，2009年，pp。 132–140。 [24] M. R. Chmielewski和J. W. Grzymala-Busse，“全球疾病 连续属性作为机器学习的预处理，”
International Journal of Approximate Reasoning，第15卷，第4号，第319–331、1996页。[25] G. K. Singh和S. Minz，“使用聚类和粗糙集理论进行离散化，in *Proceedings of the 17th International Con-*
Proceedings of the 17th International Con-
ference on Computer Theory and Applications (ICCTA)
ference on Computer Theory and Applications (ICCTA)
ference on Computer Theory and Applications (ICCTA)，li li li。 Wong 和Y. Wang，“连续数据的全局最佳算法”，*Intelligent Data Analysis*，第8卷，第151-170页。 1993。[28] J. Catlett，“在 *European Working Session on Learning (EWSL)*中将连续属性变成有序的属性”，Ser。 *Analysis*，第1卷，第1-4页，第157-179页，1997年。离散化，” *Knowledge and Information Systems*，卷。 5，pp。162–182，200 3。[32] -“重新审视有效的多层命中：optima-preseraving contition候选者”，” *Data Mining and Knowledge Discovery*，第1卷，第1卷。 8，pp。97–126，2004。[33] L. Breiman，J。Friedman，C。J. Stone和R. A. Olshen，*Classifi- cation and Regression Trees*。 Chapman and Hall/CRC，1 984年。[34] S. R. Gaddam，V。V。Phoha和K. S. Balagani，“ K-Means + ID3：一种通过cascading K- casecading K-含义K-的新颖方法，是指k-含义聚类和ID3决策树学习方法，” 19，pp。345–354，2007。[35] H.-W. 胡，Y.-L。 Chen和K. Tang，“用连续的La-Bel构建决策树的动态离散方法”，*IEEE Transactions on Knowledge and Data Engineering*，第1卷。 21 ，否。 11，第1505–1514页，2009年。[36] H. Ishibuchi，T。Yamamoto 和T. Nakashima，“模糊数据挖掘：模糊离散的影响”，in *IEEE International Conference on Data Mining (ICDM)*，2001年，2001 年，第241-241-248页。 [37] A. Roy和S. K. Pal，“粗糙集体分类器的特征空间的模糊分散化”，*Pattern Recognition Letters*，第1卷。第24页，第895–902页，2003年。[38] D. Janssens，T。Brijs，K。Vanhoof和G. We ts，“评估基于成本的离散化与基于熵和错误的离散化的绩效，” *Computers & Operations Research*，第1卷。 33，不。 11，第3107–3123 页，2006年。[39] H. He和E. A. Garcia，“从不平衡数据中学习”，*IEEE Transactions on Knowledge and Data Engineering*，第1卷。 21，否。 9，p p。1263–1284，2009。[40] Y. Sun，A。K. C. Wong和M. S. Kamel，“失败数据的分类：评论”，*International Journal of Pattern Recognition*} *and Artificial Intelligence*，第1卷。 23，否。 4，pp。687–719，2009。[4 1] A. Bondu，M。Boulle和V. Lemaire，“非参数半监督的离散方法”，*Knowledge and Information Knowledge and Information Systems*，第1卷 。 24，第35-57页，2010年。[42] F. Berzal，J.-C。 Cubero，N。Mar ´n和 D. S ´Anchez，“具有数值属性的多向决策树”，*Information Sciences*，第1卷。 165，第73–90页，2004年。[43] W.-H. Au，K。C. C. Chan和A. K. C. Wong，“分区分类的连续属性的模糊方法”，” *IEEE Transactions on Knowledge Data Engineering*，vol。 18，不。 5，pp。7 15–719，2006。[44] S. Mehta，S。Parthasarathy和H. Yang，“迈向维护离散化的无效相关性”，*IEEE Transactions on*} *Knowledge and Data Engineering*，第1卷。 17，pp。1174–1185，2005。 [45] S. D. Bay，“设置采矿的多元离散化”，*Knowledge Information Systems*，第1卷。 3，第491–512，2001页。[46] M. N. M. Ga rc ´a，J。P。Lucas，V。F。F. L. Batista和M. J. P. Mart ´n，“多变量分配为稀疏数据应用程序的多变量分配，
(HAIS)，2010年，第104-111页。本文在本期刊的未来问题中被接受，但尚未得到充分编辑。内容可能在最终出版之前发生变化。

[47] S. Ferrandiz和M. Boull Ɛ, “通过重新指示的图形监督两部分进行多变量离散化”, in *Proceedings of the 4th Conference on Machine Learning and Data Mining (MLDM)*, 2005年, 第253-264页。 [48] P. Yang, J.-S。李和Y.-X. Huang, “HDD：一种基于超立方体的基于离散化的算法”, *International Journal of Systems Science*, vol。 42, 不。 4, 第557–566页, 2011年。[49] R.-P。李和Z.-O。 Wang, “基于熵的分类方法, 用于分类规则, 不一致的检查”, 在*Proceedings of the First International Conference on Machine Learning and Cybernetics (ICMLC)*, 2002年, 第243–246页中。 [50] C.-H. Lee, “基于Hellinger的离散化方法, 用于分类学习中的数字属性”, *Knowledge-Based Systems*, 第1卷。 20, pp。419–425, 2007。[51] F. J. Ruiz, C。Angulo和N. Agell, “IDD：基于监督的间隔距离的离散方法”, *IEEE Transactions on Knowledge and Data Engineering*, 第1卷。 20, 没有。 9, 第1230–1238, 2008页。[52] J. Y. Ching, A。K. C. Wong和K. C. C. Chan, “从连续和混合模式数据中归因于class的离散化, ” *IEEE Transactions on Pattern Analysis and Machine} Intelligence*, 第1卷, 第1卷。第17页, 第641–651页, 1995年。[53] J. L. Flores, I。Inza和Larra, “通过估计分布算法的估计, 包装器离散化”, *Intelligent Data Analysis*, vol。 11, 不。 5, 第525–545页, 2007年。[54] D. A. Zighed, S。rabas ƐEDA和R.Rakotomalala, “Fusinter：一种离散连续属性的方法, ” *International International Journal of Uncertainty, Fuzziness Knowledge-Based Systems*, vol。 6, pp。307–326, 1998。[55] R. Jin, Y。Breitbart和C. Muoh, “数据离散化结算”, *Knowledge and Information Systems*, 第1卷。 19, 第1-29页, 2009年。[56] K. M. Ho和P. D. Scott, “Zeta：连续变量的全局方法的全局方法”, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining (KDD)*, 1997年, 第191-194页。 [57] L. A. Kurgan和K. J. Cios, “CAIM离散算法”, *IEEE Transactions on Knowledge and Data Engineering*, 第1卷。 16, 不。 2, 第145–153页, 2004年。[58] C.-J. Tsai, C.-I。 Lee和W.-P。 Yang, “基于类 - attribute contringency系数的离散化算法”, *Information Sci- ences*, vol。 178, pp。714–731, 2008。[59] D. Ventura和T. R. Martinez, “Brace：连续价值数据离散化的范式”, 在 *Proceedings of the Seventh Annual Florida AI Research Symposium (FLAIRS)*中, 1994年, 第117-121页。 [60] M. J. Pazzani, “贝叶斯分类器中数字属性离散化的迭代改进方法”, 在 *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD)*中, 1995年, 第228-233页。 [61] A. K. C. Wong和D. K. Y. Chiu, “从不完整的混合模式数据中综合统计知识”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 第1卷。 9, 第796–805页, 1987年。[62] M. Vannucci和V. Colla, “通过SOM挖掘的持续特征有意义离散化”, in *Proocedings of the 12th European Symposium on Artificial Neural Networks (ESANN)*, 2004年, 2004年, 第489-494页。 [63] P. A. Chou, “分类和恢复树的最佳分区”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 第1卷。 13, pp。340–354, 1991。[64] R. Butterworth, D。A. Simovici, G。S. Santos和L. Ohno-Machado, “用于监督离散化的贪婪算法”, *Journal of Biomedical Informatics*, 第1卷。 37, 第285–292页, 2004年。[65] C. Chan, C。Batur和A. Srinivasan, “在基于规则的动态系统模型中确定Quantizatization间隔”, *Proceedings of the Conference on Systems and Man and and Proceedings of the Conference on Systems and Man and and Cybernetics*, 1991, 1991, 第1719-1723页。 [66] M. Boulle, “Khiops：一种连续属性的统计离散方法”, *Machine Learning*, 第1卷。 55, 第53-69页, 2004年。[67] C.-T。苏和J.-H. hsu, “用于分离实际价值属性的扩展CHI2算法”, *IEEE Transactions on Knowledge and Data Engineering*, vol。 17, pp。437–441, 2005。[68] X. Liu和H. Wang, “基于异质性标准的离散化算法”, *IEEE Transactions on Knowledge and Data Engineering*, 第1卷。 17, 第1166–1173页, 2005年。[69] D. Ventura和T. R. Martinez, “经验比较 *Proceedings of the 10th International*中的离散方法” *Symposium on Computer and Information Sciences (ISCIS)*, 1995年, 第443–450页。 [70] M. Wu, X.-C。 Huang, X。Luo和P.-L。 Yan, “基于差异可能集理论的离散算法”, *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics (ICMLC)*, 2005年, 第1752–1755页。 [71] I. Kononenko and M. R. Sikonja, “Discretization of continuous attributes using relieff,” in *Proceedings of Elektrotehnika in Racu- nalnika Konferenca (ERK)*, 1995. [72] S. Chao and Y. Li, “Multivariate interdependent discretization for continuous attribute,” in *Proceedings of the Third International Conference on Information Technology and Applications (ICITA) Vol- ume 2*, 2005, pp. 167–172。 [73] *Machine Learning (ICML)*, 1995年, 第456–463页。 [75] [76） [77] N. Friedman和M. Goldszmidt, “在学习贝叶斯网络的同时进行持续的贡献”, 在*Proceedings of the 13th International Conference on Machine Learning (ICML)*中, 1996年, 第157-165页。 [78） 19, 第17-28页, 2007年。[79] J. Cerquides和R. L. D. Mantaras, “提案和经验比较基于可行距离的离散方法”, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)*, 1997年, 第139-142页。 [80] R. Subramonian, R。Venkata和J. Chen, “属性离散化的视觉交互式框架”, *Proceedings of the First International Conference on Knowledge Discovery and Data Mining International Conference on Knowledge Discovery and Data Mining (KDD)*, 1997年, 第82-88页。 [81] D. A. li, “一种用于离散连续属性的新颖CHI2算法”, *Proceedings of the 10th Asia-Pacific web conference on Progress in WWW research and development*, ser。 Apweb, 2008年, 第560-571页。 [83] S. J. Hong, “将上下文信息用于特征排名和离散化”, *IEEE Transactions on Knowledge and Data Engineering*, vol。 9, 第718–730页, 1997年。[84] L. Gonz ´Alez-Abril, F。J. Cuberos, F。Velasco, F。Velasco和J. A. Ortega, “Ameva：自主离散算法, ” 36, pp。5327–5332, 2009。[85] K. Wang和B. Liu, “多个属性的同时离散化”, *Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, 1998年, 第250-259页。 [86] P. Berka和I. Bruha, “各种分解程序的经验比较”, *International Journal of Pattern Recognition and Artificial Intelligence*, 第1卷。 12, 不。 7, 第1017–1032、1998页。[87] J. W. Grzymala-Busse, “基于熵离散的多重扫描策略”, *Proceedings of the 18th International Symposium on Foundations of Intelligent Systems*, ser。 Ismis, 2009年, 第25-34页。 [88] P. Perner和S. trautzsch, “决策树学习的多间隔离散化方法”, *Advances in Pattern Recognition, Joint IAPR International Workshops SSPR 98 and SPR 98*, 1998年, 第475–482页。 [89] S. Wang, F。Min, Z。Wang和T. Cao, “offd：幼稚贝叶斯分类的最佳频率离散化”, in

Proceedings of the 5th International Conference on Advanced Data Mining and Applications，Ser。 ADMA，2009年，第704–712页。 [90] S. Monti和G. F. Cooper，“一种从混合数据中学习贝叶斯网络的多元离散方法”，在*Proceedings on Uncertainty in Artificial Intelligence (UAI)*，1998年，第404-413页中。 [91] J. Gama，L。Torgo和C. Soares，“连续属性的动态离散化”，*Proceedings of the 6th Ibero-American Conference on AI: Progress in Artificial Intelligence*，ser。 Iberamia，1998年，第160-169页。 [92] P. Pongaksorn，T。Rakthanmanon和K. Waiyamai，“ DCR：使用类信息的离散化来减少intervals的数量”，in *Proceedings of the International Conference on Quality issues, measures of interestingness and evaluation of data mining issues, measures of interestingness and evaluation of data mining model (QIMIE)*，2009年，2009年，第17-28页。 [93] S. Monti和G. Cooper，“用于多变量离散化的潜在变量模型”，在*Proceedings of the Seventh International Workshop on AI & Statistics (Uncertainty)*，1999年。[94] H. [95] A. An和N. Cercone，“学习分类规则的连续属性的离散化”，Ser。人工智能中的讲义，第1卷。 1574，1999，第509–514页。 [96] S. Jiang和W. Yu，“无监督特征离散化的局部密度方法”，*Proceedings of the 5th International Conference on Advanced Data Mining and Applications*，ser。 ADMA，2009年，第512-519页。 [97] E. J. Clarke和B. A. Barton，“贝叶斯信念网络连续变量的熵和MDL离散化”，*International Journal of Intelligent Systems*，第1卷。 15，第61–92页，2000年。[98] M.-C.。 Ludl和G. Widmer，“关联规则挖掘的相对无监督的分散设备”，in *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge ropean Conference on Principles of Data Mining and Knowledge Discovery*，ser。欧洲第四届欧洲原则和知识实践会议（PKDD）（PKDD），2000年，第148-158页。 [99] A. Berrado和G. C. Runger，“与随机森林的混合数据中有监督的多变量分解”，*ACS/IEEE International Conference on Computer Systems and Applications International Conference on Computer Systems and Applications (ICCSA)*，2009年，第211-217页。 [100] F. Jiang，Z. Zhao和Y. Ge，“用于粗糙集的监督和多元离散算法”，*Proceedings of the 5th international conference on Rough set and knowledge technology*，ser。 RSKT，2010年，第596–603页。 [101] J. W. Grzymala-Busse和J. Stefanowski，“规则诱导的三种离散方法”，*International Journal of Intelligent Systems*，vol。 16，不。 1，pp。29–38，2001。[102] F. E. H. Tay和L. Shen，“用于解散的修改CHI2算法”，*IEEE Transactions on Knowledge and Data Engineering*，第1卷。 14，pp。666–670，2002。[103] W.-L. Li，R.-H。 Yu和X.-Z. Wang，“决策树生成中连续价值属性的离散化”，*Proocedings of the Second International Conference on Machine Learning and Cybernetics (ICMLC)*，2010年，第194-198页。 [104] F. Muhlenbach和R. Rakotomalala，“多变量监督离散化，邻居图方法”，*Proceedings of the 2002 IEEE International Conference on Data Mining*，ser。 ICDM，2002年，第314-320页。 [105] W. [106] A. Gupta，K。G。Mehrotra和C. Mohan，“基于聚类的监督学习离散化”，*Statistics & Probability Letters*，第1卷。 80，不。 9-10，第816–824页，2010年。[107] R. Gir´Aldez，J。Aguilar-Ruiz，J。Riquelme，J。Riquelme，F。Ferrer-Troyano和D. Rodr´guez-Baena，“与决策规则生成”，“ p } *Frontiers in Artificial Intelligence and Applications Frontiers in Artificial Intelligence and Applications Frontiers in Artificial Intelligence and Applications Frontiers in Artificial Intelligence and Applications*。 275–279。 [108） [109] J.-H. dai and y.-X. li，“基于粗糙集理论的离散化研究”，在*Proceedings of the First International Conference on Machine Learning and Cybernetics (ICMLC)*，2002年，第1371–1373页中。 [110] L. nemmiche-alachaher，“数据拆卸的上下文方法”，*Proceedings of the International Multi-Conference on Computing in the Global Information Technology (ICCGI)*，2010年，第35-40页。 [111] C.-W。 Chen，Z.-G。李，S.-y。 Qiao和S.-P。 Wen，“基于遗传算法的粗糙集合研究的研究”，*Proceedings of the Second International Conference on Machine Learning and Cybernetics (ICMLC) and Cybernetics (ICMLC)*，2003年，第1430–1434页。 [112] J.-H. Dai，“一种用于决策系统离散化的遗传算法”，在*Proceedings of the Third International Conference on Machine Learning and Cybernetics (ICMLC)*，2004年，第1319–1323页中。 [113] S. A. MacSkassy，H。Hirsh，A。Banerjee和A. A. Dayanik，“使用文本分类器进行数值分类”，in *Proceedings Proceedings of the 17th International Joint Conference on Artificial Intelligence - of the 17th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI)*，2001年，2001年，第885-890页。 [114] J. Alcalá-Fdez，L. Sánchez，S. Garc´a，M. J. del Jesus，S. Ventura，J. M. Garrell，J. Otero，C. Romero，J. Bacardit，V. M. Rivas，J. C. Fernández，and F. Herrera，“KEEL: a software tool to assess数据挖掘问题的进化算法，” *Soft Computing*，vol。 13，否。 3，第307–318，2009页。[115] J. Alcal´a-fdez，A。Fern´Andez，J。Luengo，J。Derrac，J。Derrac，S。Garc´a，L。S {算法和实验分析框架，” *Journal of Multiple-Valued Logic and Soft Computing*，第1卷。 17，不。 2-3，第255–287页，2011年。[116] A. Frank和A. Asuncion，“ UCI机器学习存储库”，2010年。[在线]。可用：http：//archive.ics.uci.edu/ml [117] K. J. Cios，L。A. Kurgan和S. Dick，“高度可扩展且可靠的规则学习者：绩效评估和比较”，*IEEE Transactions on Systems, Man, and Cybernetics, Part B*，第1卷。 36，第32-53页，2006年。[118] D. R. Wilson和T. R. Martinez，“基于实例的学习算法的还原技术”，*Machine Learning*，第1卷。 38，不。 3，第257–286页，2000年。[119] D. W. Aha，ed。，*Lazy Learning*。 Springer，2010年。[120] E. K. Garcia，S。Feldman，M。R。Gupta和S. Srivastava，“完全懒惰的学习”，*IEEE Transactions on Knowledge and Data Engineering*，第1卷。 22，pp。1274–1285，2010。[121] R. Rastogi和K. Shim，“公共：整合建筑物和修剪的决策树分类器”，*Data Mining and Knowledge Discovery*，第1卷。 4，pp。315–344，2000。[122] W. W. Cohen，“快速有效的规则诱导”，*Proceedings of the Twelfth International Conference on Machine Learning (ICML)*，1995年，第115-123页。 [123] J. A. Cohen，“名义量表的协议系数”，*Educational and Psychological Measurement*，第37–46页，1960年。[124] R. C. Prati，G。E. A. A. P. A. Batista和M. M. C. Monard，“一项针对分类的预测性能评估的图形方法调查，” 125} *IEEE Transaction on Knowledge and Data Engineering, IEEE Transaction on Knowledge and Data Engineering, IEEE Transaction on Knowledge and Data Engineering,* {{{{{{{a。本·戴维（Ben-David），“很多随机性都隐藏在准确性中”，*Engineering Applications of Artificial Intelligence*，第1卷。 20，pp。875–885，2007。[126] J. dem ˇSAR，“分类器对多个数据集的统计比较”，*Journal of Machine Learning Research*，vol。 7，第1-30页，2006年。[127] S. garc´a和F. herrera，“分类器的统计分类器上的分类器上的统计量对多个数据集的统计分解”，以进行所有成对比较，” *Journal of Machine Learning Research*，第1卷，第1卷。 9，第2677–2694，2008页。[128] S. garc´a，A。Fern´Andez，J。Luengo和F. Herrera，“高级非参与测试，用于在计算智能和数据挖掘的计算智能和数据矿井分析中进行多个实验设计的多个实验的比较：实验分析：VORES of POWER的实验：V86。 180，没有。 10，第2044–2064页，2010年。[129] F. Wilcoxon，“通过排名方法进行的个人比较”，*Biometrics*，第1卷。 1，第80–83页，1945年。

Salvador Garc ´a获得了硕士学位和博士分别于2004年和2008年获得西班牙格拉纳达大学的计算机科学博士学位。

他目前是JA ´EN，JA ´EN的计算机科学系助理教授，西班牙。他在国际期刊上发表了25篇论文。他在不同的数据挖掘主题上共同编辑了国际期刊的两个特殊问题。他的搜索兴趣包括数据挖掘，数据降低 - tion，数据复杂性，不平衡学习，半监督学习，统计推断和进化算法。

弗朗西斯科·埃雷拉（Francisco Herrera）获得了硕士学位在1988年的数学出版社和博士学位1991年的数学博士学位，均来自西班牙格拉纳达大学。

他目前是格拉纳达大学计算机科学和艺术智力的教授。他在国际期刊上发表了200多篇论文。他是“遗传模糊系统：进化调整和学习模糊知识基础”一书的合着者（世界学业 - TIFIF，2001年）。

他目前担任国际杂志“人工智能进展”（Springer）的主编（Springer），并担任《软计算》杂志（进化和生物启发算法）和国际计算情报系统杂志的区域编辑。他充当期刊的相关编辑：关于模糊系统，信息系统，模糊系统进步和国际应用元启发式计算杂志的IEEE交易；他是几个期刊编辑委员会的成员，包括：模糊集和系统，应用智能，知识和信息系统，信息融合，进化智能，国际混合智能系统杂志，模因计算，群和进化计算。

Juli ´uengo获得了硕士学位科学和博士学位2006年和2011年从西班牙格拉纳达大学（Granada）分别出发。他目前是西班牙布尔戈斯大学大学土木工程系的助理教授。他的研究包括机器学习和数据挖掘，知识发现和数据挖掘中的数据准备，缺失值，数据复杂性和模糊系统。

他获得了以下荣誉和奖项：ECCAI研究员2009年，2010年西班牙国家计算机科学奖Aritmel授予“西班牙计算机科学工程师”，以及国际Cajastur“ Mamdani”软计算奖（第四版，2010年）。

他当前的研究兴趣包括用单词和删除制作，数据挖掘，数据制备，实例选择，基于模糊规则的系统，遗传模糊系统，基于进化算法的知识提取，模因算法，模因算法和遗传算法的知识。

jos ´e antonio s ´aez

获得了他的硕士学位2009年，西班牙格拉纳达大学格拉纳达大学的计算机科学博士学位。他目前是博士学位。西班牙格拉纳达大学的计算机科学与艺术情报系的学生。他的研究兴趣包括数据最低，数据预处理，基于模糊规则的系统和不平衡的学习。

维多利亚L ´Opez获得了他的硕士学位2009年，西班牙格拉纳达大学的格拉纳达大学科学学士学位。她目前是博士学位。西班牙格拉纳达大学的计算机科学与艺术情报系的学生。她的研究包括数据挖掘，失败域中的分类，模糊规则学习和进化算法。