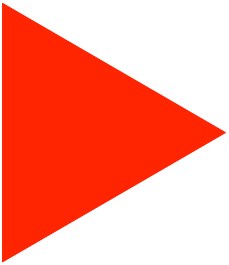


**大型相异数据的证据聚类**

Thierry Denoeux, Songsak Sriboonchitta, Orakanya Kanjanatarakul

**引用此版本:**



Thierry Denoeux, Songsak Sriboonchitta, Orakanya Kanjanatarakul. Evidential clustering of large dissimilarity data. Knowledge-Based Systems, Elsevier, 2016, 106, pp.179-195.

<10.1016/j.knosys.2016.05.043>. <hal-01324491>

**HAL Id: hal-01324491**

[**https://hal.archives-ouvertes.fr/hal-01324491**](https://hal.archives-ouvertes.fr/hal-01324491)

Submitted on 1 Jun 2016

**HAL**是一个用于存放和传播无论是否出版的科学研究文件的多学科的开放存取档案馆，这些文件可能来自法国或国外的教学和研究机构，也可能来自公共或私人研究中心。

多学科开放档案馆HAL，用于存放和传播法国或外国教育和研究机构，公共或私人实验室出版或未出版的科研水平文件。

大型相异数据的证据聚类

Thierry Denœux∗[[1]](#footnote-1), Songsak Sriboonchitta† and

Orakanya Kanjanatarakul‡

∗ Sorbonne Universit´es

Universit´e de Technologie de Compi`egne, CNRS,

UMR 7253 Heudiasyc, France

1. Faculty of Economics, Chiang Mai University, Thailand
2. Faculty of Management Sciences,

Chiang Mai Rajabhat University, Thailand

May 24, 2016

# 摘要

在证据聚类中,对象到聚类的成员资格被认为是不确定的,并且由Dempster-Shafer质量函数表示，形成一个信用分区。EVCLUS算法以这样的方式构造信用分区，即对象之间的较大不相似性对应于相关联的质量函数之间的较高程度的冲突。在本文中，我们对EVCLUS进行了一些改进，使其适用于非常大的相异性数据。首先，原始EVCLUS算法中基于梯度的优化过程被更快速的迭代行式二次规划方法所取代。 其次，我们证明EVCLUS只能提供不同的随机样本，将时间和空间复杂度从二次变为大致线性。最后，我们引入了一个两步法构建信用分区，将质量分配给选定的簇对，使得算法输出的信息量比原始EVCLUS的信息更多，同时保持对大量簇的管理。

**关键词：**Dempster-Shafer理论，证据理论，信度函数，无监督学习，信用分区，关系数据，邻近数据，成对数据。

# 引言

将数据聚类成组是数据挖掘和机器学习的基本任务之一。 可以根据它们可以处理的输入数据以及它们产生的输出来区分聚类算法。

通常，考虑两类输入数据：属性（矢量）数据和不相似（接近，关系，成对）数据。在前一种情况下，每个对象由数字或分类属性的向量描述。在后者中，数据采用对象之间的不相似矩阵的形式。通过选择合适的距离，可以容易地将属性数据转换为相异性数据。 逆变换（从相异性到属性数据）通常更难，除非是在度量不相似的特殊情况下，即在潜在空间中的矢量之间的精确欧几里德距离的不相似性，在实践中不那么频繁的情况。而找到一组对象的属性表示，使得对象之间的距离接近给定的相异度矩阵又通常是一项困难的任务（称为多维缩放 - MDS），这需要解决大规模的非线性优化问题[3,4]。大多数聚类算法（例如c-means算法及其众多变体）旨在处理属性数据。只有比较少数量的算法（称为关系聚类方法）可以直接处理相异性数据[9–11]。

至于聚类输出，我们可以区分为用于查找对象分区的分区聚类和查找嵌套分区序列的分层聚类。多年来，分区聚类的概念已经扩展到几个重要的变体，包括模糊[2]和可能性[16]聚类，以及最近的粗糙[20,27]和证据[7,25]聚类。与经典（硬）分区聚类相反，其中每个对象都被明确地分配并且完全确定地指向单个聚类，这些变体允许对象分配给聚类时的模糊性，不确定性或怀疑性。出于这个原因，相对于经典的“硬”聚类，它们被称为“软”聚类方法[28]。在软聚类范例中，证据聚类使用Dempster-Shafer质量函数[30]描述每个对象的成员资格对集群的不确定性，而该函数为每个聚类子集分配质量。这是对数据集的聚类结构的丰富且翔实的描述，并且可以显示作为特殊情况的硬、模糊和粗糙分区。最近，证据聚类已成功应用于各种领域，如机器预后[29]，医学图像处理[17,24]和社交网络分析[34]。 在监督分类中也已经开发了类似的想法（参见，例如，[18,21,22]）。

在[7]中，我们中的一位（第一作者）介绍了EVCLUS，这是一种用于处理（非必需度量）相异数据的证据聚类算法。EVCLUS基于这样的自然假设：当两个对象更相似时，属于同一群集的两个对象的合理性更高。这种假设转化为搜索最小化成本函数的信用分区。 EVCLUS的变体允许人们以成对约束的形式使用先验知识，后来在[1]中引入。

EVCLUS算法有几个优点，首先它在概念上很简单，并且它可以处理非度量不相似数据（甚至以顺序尺度表示），它还被证明在许多数据集上胜过一些关于状态的关系聚类技术[7]。从弊端来看，EVCLUS的主要缺点是其计算复杂性，与其他关系聚类算法一样，它需要存储整个不相似矩阵，因此，空间复杂度为O（n2），其中n是对象的数量，而这就排除了对包含超过几千个对象的数据集的应用。此外，在EVCLUS算法中实现的基于梯度的优化过程的每次迭代都需要O（f3n2）的运算，其中f是质量函数的焦点集的数量，即正在考虑的集群的子集的数量。而在最坏的情况下，f = 2c，其中c是簇的数量。 为了使该方法即使对于c的中等值也可用，我们需要限制质量函数的形式，以便仅将质量分配给大小为0,1或c的焦点集，而这阻止了我们充分利用其潜在的一般性方法。

在本文中，我们提出了对EVCLUS算法的一些改进，使其适用于非常大的数据集。 这些改进有三个方面。首先，原始EVCLUS算法中基于梯度的优化过程被[31]中提出的更快速迭代行式二次规划方法的改编所取代。其次我们证明了EVCLUS并不需要提供整个相异矩阵，从而将时间和空间复杂度从二次变为大致线性。 最后，我们引入了一个两步法构建信用分区，将质量分配给选定的簇对，使得算法输出的信息量比原始EVCLUS的信息更多，同时也保持对大量簇的管理。

本文的其余部分安排如下。关于信度函数，证据聚类和EVCUS算法的背景将在第2节中首先回顾，新的优化过程将在第3节中进行描述和评估，然后将分别在第4节和第5节中描述EVCLUS的改进，使其适用于具有大量对象和大量簇的问题，最后第6节本文结束。

# 背景

在本节中，首先会在2.1节中提供关于Dempster-Shafer理论的简要提醒，然后在2.2节中回顾信用分区和其相关的必要概念，EVCLUS算法将在2.3节中介绍。

## 质量函数

设Ω= {ω1，...，ωc}是一个有限集，表示对某个问题Q的可能答案，其中有且只有一个为真。真正的答案用ω表示。质量函数Ω是从功率设置为2Ω到[0,1]的映射，使得

X *m*(*A*) = 1*.* (1)

*A*⊆Ω

每个数m（A）代表与命题ω∈A相关的支持度，并且不再代表具体命题[30]。 Ω的使得mi（A）> 0的子集A被称为m的焦点集。 质量函数m可以是：

* 如果∅不是焦点集，是归一化函数;
* 如果只有一个焦点集，是逻辑函数;
* 如果它的焦点集是独立的，是贝叶斯方法
* 如果它只有一个焦点集，并且该焦点集是单例，同时是逻辑函数和贝叶斯
* 如果其焦点集是嵌套的，是兼并的。

对于每个质量函数m，我们可以将信度函数和似然函数分别定义为从2Ω到[0,1]，如下所示，

*Bel*(*A*) = X *m*(*B*) (2a)

∅6=*B*⊆*A*

*Pl*(*A*) = X *m*(*B*)*,* (2b)

*B*∩*A*6=∅

对于所有A⊆Ω，这两个函数都可以通过关系式P1（A）= Bel（Ω）-Bel（A）连接起来。对于所有的A⊆Ω，数量Bel（A）是已知证明有多少支持命题“ω∈A”量度。相反，Bel（Ω） - Pl（A）= Bel（A）是支持互补假设A的程度的度量，因此Pl（A）可以被视为不对A支持的度量。pl：Ω→[0,1]将Ω的每个元素ω映射到其合理性pl（ω）= P1（{ω}），称为与m相关的轮廓函数。如果m是贝叶斯，则Bel = P1，并且该函数是概率测度; 因此，轮廓函数是通常的概率质量函数，即对于所有A⊆Ω，*Bel*(*A*) = *Pl*(*A*) = P*ω*∈*A pl*(*ω*)。 如果m是辅音，则P1是可能性度量，即对于所有A，B⊆Ω，具有P1（A∪B）= max（P1（A），P1（B）），并且Bel是双重必要性度量; 而对应于所有A⊆Ω，pl是相应的可能性分布，即*Pl*(*A*) = max*ω*∈*A pl*(*ω*)。辅音质量函数可以从其轮廓函数中唯一地恢复。

1. Corresponding author. Phone: +33 344 234 496, fax: +33 344234477, email: tdenoeux@utc.fr. [↑](#footnote-ref-1)