计算机科学2004Vol.31№.2

# 抽样在数据挖掘中的应用研究*)

张春阳　周继恩　钱　权　蔡庆生
(中国科学技术大学计算机系　合肥230027)

**摘　要**　大规模数据集是数据挖掘高效实现的障碍。抽样是统计学中一种常用的调查方法，作为克服该障碍的方法，抽样被引入数据挖掘中。在国外，抽样在数据挖掘中的应用研究已比较广泛，而国内相关研究很少。本文在总结现有相关工作的基础上，系统介绍了数据挖掘中抽样的应用及其相关问题。相信抽样在数据挖掘中的应用研究将推动数据挖掘的发展。

**关键词**　抽样，数据挖掘，策略

## Research of Sampling's Application in Data Mining

ZHANG Chun-Yang　ZHOU Ji-En　Qian Quan　CAI Qing-Sheng
(Comp. Department of Univ. of Sci. and Tech. of China, HeFei 230027)

**Abstract**　Large data sets are becoming obstacles for efficient data mining. Survey sampling is used in statistics commonly. In data mining, we use sampling to overcome the large data sets. Many works have been done about sampling's application in data mining. Based on the summarization of current relational works, we introduce systemically the sampling's application in data mining.

**Keywords**　Sampling, Data mining, Strategy

## 1　引言

抽样调查是调查的最基本形式之一，它按照一定的程序，从全体调查对象中抽取一部分进行调查，然后根据样本数据对总体目标进行估计。20世纪20年代起抽样调查的基本理论逐步形成。抽样调查因其节约时间、节约花费、高正确性和应用领域广等特性，被广泛应用于社会调查的各个领域中[1]。

数据挖掘是从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘有趣知识的过程。数据挖掘技术自兴起以来，一直是研究的热点问题，现已有大量的实现算法。但随着数据处理技术的飞速发展，需要处理的数据规模越来越大，且由于数据分析内部的复杂性，现有算法在进行大规模数据分析时还需要消耗大量的时间和物理资源。如何减少大规模数据分析所消耗的资源?在提高数据挖掘算法效率的同时，我们很自然地会引进统计中的抽样思想，首先对大规模数据集的部分数据进行数据分析，然后根据相应的结果进行下一步处理，以提高数据分析的效率，这是一种行之有效的方法。

## 2　统计学中的抽样调查简介[1,2]

### 2.1　抽样调查的基本概念及术语

总体(population)：所研究(调查)对象的全体。

样本(sample)：按某种方法从总体中抽取其中一部分个体进行调查，这部分个体就称为样本。

抽样方法：样本的抽取方法。抽样可分为概率抽样(probability sampling)和非概率抽样(non-probability sampling)两类。概率抽样也称为随机抽样(random sampling)。概率抽样是一种从总体中按一定概率获取样本的方法。概率抽样的优点是能够保证样本的代表性，避免人为的干扰和偏差。它还能对由于抽样引起的误差—抽样误差进行估计。因此采用概率抽样可以获得估计的精确。鉴于这两个原因，概率抽样是最科学、应用最广泛的一种抽样方法。

抽样单元(sampling unit)：为使概率抽样能够实施，同时也为了具体抽样的便利，通常将总体划分成互不重迭的若干部分，每个部分称为一个抽样单元。

样本量：样本中包含的抽样单元的数目。

### 2.2　几种基本的抽样策略

对不同的项目应采用不同的抽样策略，最基本的抽样策略有以下五种，在实际问题中，一个具体的抽样方法大多是这五种策略的各种形式的组合。

(1)简单随机抽样(simple random sampling)：从大小为N的总体中逐个不放回地抽取n个单元组成样本，每次抽取对当时尚未入样的单元都是随机抽取的，也即都是等概率的。

(2)分层抽样(stratified sampling)：将总体中的单元按某种原则划分成若干个子总体，每个子总体称为层，在每层中独立进行简单随机抽样或其它抽样。

(3)整群抽样(cluster sampling)：总体中的每个抽样单元可以分成若干个次级单元，抽样仅对初级单元抽，若某个初级单元被抽中，则调查这个单元中的所有次级单元。

(4)多阶抽样(multi-stage sampling ,subsampling)：若初级单元内的次级单元相似程度较大，调查所有次级单元会造成很大的浪费。此时一个自然的想法是在被抽中的初级单元中再对次级单元进行抽样，这就是多阶抽样。

(5)系统抽样(systematic sampling)：总体中的抽样单元按某种次序排列，在规定的范围内随机抽取一个(或一组)初

*)本文得到国家自然科学基金重大项目(90104030)的资助。张春阳　博士研究生，研究方向为数据挖掘、入侵检测、数据仓库；周继恩　博士研究生，研究方向为数据挖掘，机器学习；钱　权　博士研究生，研究方向为入侵检测；蔡庆生　教授，博导，研究方向为人工智能。

始单元，然后按一套事先确定的规则确定其他样本单元的抽样方法。

## 3 数据挖掘中的抽样

作为一个快速发展的领域，数据挖掘的功能是从大规模数据集中发掘出重要模式和有趣规则。如果一个数据集具有海量的记录或很多属性，或者两者都有，我们则称它为大规模数据集。数据挖掘在寻找有趣知识或适合模型时，一般需要多遍扫描所分析的数据集，所以大规模数据集是数据挖掘的主要障碍。随着大规模数据集越来越普遍，数据挖掘的发展也需使其更适于处理大规模数据集。抽样，对样本数据集进行挖掘，然后推测分析出总体数据集的相关信息，则是数据挖掘克服大规模数据的方法之一。

### 3.1 抽样在数据挖掘不同领域中的应用

3.1.1 *抽样在关联规则中的应用* 将抽样方法应用于关联规则的挖掘是由 Toivonen 在文[4]中首先提出的。该文提出了一种基于抽样的关联规则发现算法，该算法仅需要对数据集完整扫描一遍就可以发现几乎所有的关联规则。该算法的基本原理是，获得一个随机的样本，在这个样本上发现关联规则，并把发现的关联规则当作总体数据集的规则，然后再用数据集中剩余的数据来验证这些规则。在一般情况下，该算法对整个数据集进行一遍扫描就可以发现所有的关联规则。

文[5]提出的 FAST(Finding Associations from Sampled Transactions)算法，也是一种基于抽样的两阶段的关联规则发现算法。在第一阶段，FAST 算法首先抽样生成数据样本，并根据数据样本快速地估计数据集中每个数据项的支持度；在第二阶段，使用估计得到的数据项的支持度，在初始的样本中对"离群"数据进行裁剪或者选择更具有代表性的数据，并形成最终样本，该样本能够正确反映整个数据集合的统计特性。最后，在这个样本数据集上实现关联规则的发现。

3.1.2 *抽样在分类中的应用* 通常的分类方法有三种：判定树、神经网络和统计学方法，这些方法都借助了抽样的思想，所以分类与抽样是密不可分的。

在判定树归纳算法的 ID3 版本中用到的"窗口"(windowing)就是一种抽样策略。窗口用来选定进行判定树训练的训练数据，它的形成分为两步：1)从所有训练数据中随机抽样生成一个初始窗口；2)在初始窗口上生成初始决策树，使用剩余训练数据验证初始决策树，并根据验证结果调整窗口直到获得满意的决策树。

分类中可以采用的抽样策略很多，Catlett 在文[6]中例举并比较了四种分类中的抽样策略。这四种策略是简单随机抽样、压缩重复、改进窗口和分层化，其中后三种为非随机抽样。压缩重复就是将每组重复的事例用一条事例来表示；改进窗口就是在窗口初始化时进行评估，并选择评估结果最好的窗口作为初始化窗口；分层化就是使用分层抽样方法，目的是尽量获得连贯的数据分布。

3.1.3 *聚类中的抽样* 抽样在聚类中的应用主要在两个方面：形成初始的簇、数据约简。

聚类中最著名与最常用的划分方法是 k-平均和 k-中心点两种方法，它们执行的第一步都是抽样获取 k 个对象作为初始簇的中心。然后再根据相应的算法执行计算，调节这些中心，直到不发生变化为止。聚类中的一些算法对初始簇的选择很敏感，所以在获得初始簇的过程中，抽样策略的选择尤为重要。

抽样在聚类中也起数据约简的作用。例如，典型的 k-中心点划分算法 PAM 对小数据集非常有效，但对大的数据集没有良好的可伸缩性，这时可以采用一个基于抽样的方法 CLARA(Clustering LARge Application)，CLARA 的主要思想是：不考虑整个数据集，随机抽样实际数据的一小部分作为数据的样本，然后用 PAM 方法从样本中选择中心点。随机抽样策略产生的样本的代表性较好，所以生成的簇亦具有代表性[7]。

### 3.2 数据挖掘中的抽样策略

3.1节介绍了抽样在数据挖掘的各种分析方法中的应用，本节总结数据挖掘中应用的各种抽样策略，并将其同统计学中相应的抽样策略进行比较。

3.2.1 *简单随机抽样* 在统计学中，简单随机抽样是所有其他抽样方法的基础，但它在大规模的抽样调查中，很少被单独采用。主要原因是它需要一个对全部基本单元的完整抽样框，且所得的样本单元相当分散，调查不便。而在数据挖掘技术中，因为总体的确定性，这样的限制被消除，所以简单随机抽样是数据挖掘技术中最常用的抽样策略。这种抽样方法简单易行，仅需借助于随机数生成函数并扫描一遍总体数据集就可以产生样本数据集，且生成的样本数据集映射了总体数据集的数据分布情况，它主要可以应用在以下两个方面：

1)生成用来估计初始化参数的随机样本；

2)数据约简，生成代表总体数据的样本数据，数据挖掘算法在此样本数据上挖掘结果。

3.2.2 *分层抽样* 当训练数据集中各属性的值不是均衡分布时，随机抽样策略很难产生好的学习样本，这时可以借助统计学中分层抽样的思想和方法。分层抽样可以按照属性值分布的不同将总体数据划分为若干个层，在每层中独立进行简单随机抽样或其它抽样，这样生成的训练样本将大大减少分类学习的学习复杂度。分层抽样的这一特性是由 Catlett 在文[6]中实验证明的。

3.2.3 *窗口* 是由 Quinlan 在决策树算法的 ID3版本中提出的。它的实现机制是从总体训练数据集中选择合适的训练样本，并在此样本上生成可供学习的高质量的决策树。为了实现该机制，需要维持一个可供学习的训练样本，即窗口。该算法的运行流程如下：

1)初始化窗口，它是由训练数据集的一个小的随机训练样本构成；

2)学习算法在初始窗口上训练生成决策树；

3)用样本以外的训练数据评估决策树；

4)如果决策树没有达到评估要求，则在窗口添加更多的训练例，并循环执行2-4，直到生成的决策树达到评估要求。

由以上流程可以看出初始的窗口也是由随机抽样生成的，但窗口同随机抽样的区别是窗口是增量式的，是一种自适应的抽样方法，也就是说它通过对当前样本的评价来确定是否要停止抽样或将哪些数据单元加入样本中。

3.2.4 *密度偏差抽样*(Density biased sampling) 是在数据挖掘中产生的一种新的抽样策略[8]。它的核心思想是依据数据分布的密度情况来生成样本，以实现数据的约简。密度偏差抽样的特性[9]如下：

1)适应性：根据所选数据挖掘算法的不同，我们可以选择抽样的核心是高密度区域还是低密度区域。如果需要在具有大量噪音的数据中发现聚类，或者发现聚类中的统计信息，我们可以对高密度区域抽样。如果要发现所有的聚类，则我们既

要对高密度区域抽样，又要对低密度区域抽样。为了识别离群数据，我们可以对极低密度区域进行抽样。

2)效果：作为一种数据约简技术，密度偏差抽样比随机抽样具有更好的约简效果。

3)效率：该方法首先需要获得数据分布的密度函数，这需要完整扫描一遍总体数据集，然后需要1-2遍扫描总体数据集来生成样本。

密度偏差抽样的难点是数据分布密度函数的获得。在文[8]通过哈希(Hashing)方法将所有数据分配入不同的聚合内，将每个聚合内的数据的个数作为数据分布的密度，并遵循以下原则抽样：

1)在同一个聚合内，每个点被选中的可能性是相等的；

2)样本保留数据分布的密度信息；

3)抽样的偏差依据聚合的大小；

4)样本量为M。

在该文中密度偏差抽样取得了很好的实验效果，而且特别需要注意的一点是简单随机抽样策略作为一个的特例被包含在密度偏差抽样策略中。

简单随机抽样、分层抽样、窗口和密度偏差抽样是数据挖掘和数据库中常用的抽样策略，其中简单随机抽样的应用最为广泛。在数据挖掘中还有很多其它的抽样策略，如：渐近抽样(Progressive sampling)[10]、动态抽样(Dynamic sampling)[11]、不确定抽样(Uncertainty sampling)[12]和重抽样(Resampling)[13]等。这些抽样策略有些引入了统计学中的思想和方法，有些仅引入了思想，有些是专门为数据挖掘设计的。无论抽样策略的来源如何，了解数据挖掘中抽样的特性是应用好抽样的前提。

### 3.3 数据挖掘中抽样的特性

数据挖掘中抽样的特性主要有以下三点：(1)并不是所有统计学中的抽样策略都可以应用在数据挖掘中；(2)抽样策略的选择，主要依据不同的数据挖掘的分析目的和数据集的不同特性；(3)因为新应用的产生，可能会有新的抽样策略的生成。

数据挖掘中的抽样同传统的抽样具有区别。第一，问题发生的起点不同。在统计学中，总体中的数据可能仅有一小部分是可以获得的，抽样在很多情况下是获得总体信息的唯一办法。而在数据挖掘中，总体数据是存在的，但困难是怎么去研究这些数据。第二，它们的目的不同。数据挖掘中的抽样主要是以下三个目的：(1)克服计算机硬件设备的限制；(2)降低学习算法的复杂度；(3)在适当降低正确性的情况下，加快学习的速度。统计中抽样的目的是推算总体的属性。第三，在样本上将进行不同的分析操作。在数据挖掘中，数据挖掘算法将用样本取代总体来获取知识；统计中，样本用来推测未知总体的有用特性。

传统的抽样方法有很多种，它们中的一部分可以以不同形式应用在数据挖掘中。有些策略，如简单随机抽样，可以被完全应用到数据挖掘中，包括它的思想和方法；而有些策略，因为数据挖掘的特殊应用要求，仅能借用其思想，譬如整群抽样。

### 3.4 抽样在数据挖掘中需要考虑的问题

在3.3节中我们介绍在数据挖掘中应用抽样主要有三个目的，但确保采用抽样能够达到这三个目需要考虑多方面的问题，以下为其中的重点问题。

**3.4.1 抽样的效率** 效率是数据挖掘算法最重要的因素之一，正确性也是。所处理的数据越多，得到结果的正确性就越高，但是同时却降低了效率；相反所处理的数据减少时，效率会增加但是降低了正确性。数据挖掘中通过抽样来获得结果，是以正确性换取效率的策略。

一般来说，一种抽样策略需要扫描几遍总体数据集的主要依据是样本的获取方式和所研究的数据的格式。对一般的关系数据库，简单随机抽样策略扫描的次数不会高于一次。然而其他抽样策略，如分层抽样，如果层的信息需要通过计算获得，则至少需要扫描两遍数据集，一遍用来生成层的信息，一遍用来生成样本。如果数据库的存储格式不支持随机访问的方式，那么简单随机抽样策略扫描总体数据库的次数将高于一次[3]。

因此在决定实施抽样以前，必须对比抽样所带来的额外开销同其所带来效率，避免既降低了数据挖掘的正确性，同时又降低了效率。

**3.4.2 样本量** 抽样后数据挖掘分析的正确性和效率是一对矛盾体，而样本量是决定抽样后数据挖掘分析的正确性和效率的重要因素之一。样本量大，正确性高，效率低；样本量小，正确性低，效率高。

因为应用的出发点不同，统计学和数据挖掘对样本量的要求也不同。在统计学中认为小样本量就满足分析的需要，譬如对100万个事例，可以依赖一个包含1000个事例的样本产生正确的统计结果。而数据挖掘中小样本量将丢失大量有用信息，很难满足分析的需要。文[14]研究了样本量在数据挖掘和统计中的作用。

多大的样本量满足学习的要求，样本量是否越大越好。总体数据集中数据分布的特性、数据挖掘算法、数据挖掘算法的正确性要求程度等因素决定此类问题的答案[4]。所以在确定样本量时要综合考虑以上因素。

**3.4.3 样本的质量** 就是样本对于总体数据集的代表性。对一个给定的数据挖掘分析方法、抽样策略和样本量，我们要通过对该样本的数据状况和抽样误差来评价[15]。

样本的数据状况主要在两个方面进行评价。第一，样本数据的覆盖程度，需要首先了解样本中数据每个属性的取值范围，然后将其同总体数据中相应属性的取值范围相比较，根据比较的结果得出样本数据的覆盖程度。第二，样本数据的分布状态，如果是随机抽样，我们希望样本数据的分布状态与总体数据的分布状态相近，而如果是偏离抽样，则我们需要验证样本是否达到偏离的要求。对这两个方面评价的综合将得到样本的数据状况。

由于用局部的样本数据来代替总体数据，因此会产生抽样误差。不同的数据挖掘方法和不同的抽样策略会产生不同抽样误差，如在聚类分析中，随机抽样会导致小簇的丢失[8]，而在关联分析中，抽样将导致部分规则的信任度的提高[5]。在统计学中，对抽样误差估计已有坚实的理论基础[1]，而在数据挖掘中，抽样误差的估计还是一个难点问题。

**总结** 本文总结了抽样在数据挖掘中的应用现状。在现有的相关工作中，抽样已在数据挖掘中得到了广泛的应用并取得了一些效果。数据挖掘中的抽样思想来源于统计学中的抽样调查。统计学中的抽样理论经过近百年的发展，已有坚实的基础。由于数据挖掘对抽样的特殊应用要求以及和统计学中抽样的区别，数据挖掘的抽样仅能借鉴统计中为数不多的理论和方法，因此抽样在数据挖掘中应用还有许多问题需要

(下转第141页)

① 图像库管理模块：负责新图像入库以及系统对图像库的连接工作，填充图像数据库的各个基本属性值，如：图像ID，创建日期，大小，关键字类别等。

② 图像特征提取模块：对新入库的图像，利用系统的特征提取算法提取图像的多种基本特征，如：颜色，形状，纹理等，在图像基本特征库中生成该图像的特征属性值序列。同时，分别生成对应于这些基本特征属性的索引结构。在此基础上，调用系统的语义特征层次模块，推理出该图像的高层语义特征，在语义分类特征库中生成其对应的特征属性值。

语义分类特征属性值可以分为以下两种：

a. 描述性语义分类关键字，如：风景，日出等。这类属性值一般由用户在图像入库或相关反馈阶段手工添加。在这种属性值间实现的是精确性匹配；

b. 一组图像基本特征的特定组合，用来反映特定的图像语义。

③ 图像查询处理模块：处理用户提交的图像查询请求。模块首先分析用户查询的查询请求，将其转换成查询特征向量后，调用图像相似性匹配算法进行匹配计算。如果待查询图像不是新图像，则直接从图像的基本特征库和语义分类特征库中调出其特征属性值序列，输入系统的图像匹配算法进行计算；否则先调用(2)提取其特征，再进行计算。

④ 用户相关反馈模块：图像相似性匹配算法完成后，需要对用户提交的相关反馈进行处理，调用相关反馈计算模型调整搜索策略，重新进行图像匹配[5]。对于相关反馈过程中新学习到或新标注的语义特征，还要写入语义分类特征库中。

(3)*数据访问层模块* 主要负责图像基本视觉特征库和语义分类特征库的管理和连接访问操作，并负责其与图像库中对应图像的一致性访问。

MVC 模型的应用使得系统模块分明，逻辑清晰；系统配置灵活，维护方便，便于今后的功能维护和扩充。同时又保证系统具有较高的整体化程度，各个子模块能够协同工作，保证了数据一致性。由于在设计中采用的是 JSP、Servlet 和 JavaBean 这一套纯 Java 的技术，系统几乎无需移植即可以运行在 Unix、Windows、Linux 等一系列主流操作系统之上，具有较强的平台无关性。此外，这种三层架构设计使得图像检索的 Web 应用能够稳定、安全地运行。

### 3.2 检索流程

基于内容的图像检索是根据用户的要求逐步求精的检索过程[5]；@IMAGE 系统对用户给出的示例图像进行特征提取，然后与图像特征库的特征进行相似匹配，查找一组候选的图像，供用户选择，如果不满意用户可给出相关反馈的要求，然后进行调整检索策略，再重新进行匹配查询，如此地缩小查询的范围，直到查询到用户满意的图像。该系统的检索处理流程如下：

(1)*图像查询要求* 用户查找一种图像时，首先给出颜色、形状或纹理等的查询特征，系统将提取该示例的特征或把查询描述为具体的特征矢量。

(2)*相似性匹配检索* 根据用户输入的图像查询要求，系统调用相应的基于图像颜色、形状、纹理的检索算法或综合检索算法与特征库中的特征进行相似匹配。

(3)*查询的候选图像* 将满足一定相似性条件的一组候选图像，按相似度大小排列后，返回给用户。

(4)*用户相关反馈和特征调整* 对系统返回的候选图像，用户可以根据要求来挑选，如得不到满意的结果，通过用户相关反馈模块给出反馈的信息，并从候选结果中选择若干示例并作"优""劣"标记，经过特征调整后，形成一个新的查询，直到得到满意的查询结果。

**结束语** 基于内容的图像检索是图像处理和识别、数据库、模式识别、多媒体和人工智能等领域的技术的综合应用，本文对基于内容的图像检索的关键技术进行了研究，主要包括：图像的特征抽取、图像数据库技术、特征空间和相似性、基于图像颜色、形状和纹理的检索算法以及综合算法[7]，并研制成功一个适用于 Web 环境的基于内容的图像检索系统。经大量的实验应用表明，其效果很好。该系统可应用于医学、工业、管理、航空、公安政法、军事和文化娱乐等领域。

## 参考文献


1 李瑜，李磊. 基于内容的图像检索方法研究. 计算机科学，1999，26(8)：6～11

2 庄越挺，潘云鹤. 基于内容的图像检索综述. 模式识别与人工智能，1999，12(2)

3 黄祥林，沈兰孙. 基于内容的图像检索技术研究. 电子学报，2002，30(7)

4 Benitez A B，Packs，Chang S E. Object-based Multimedia Content Description Schemes and Application for MPEG-7，Signal Proceessing：Image Communication，2000，16：235～269

5 Huang J，et al. Image indexing Using Color Correlogram. In：Proc. of IEEE Conf. on computer Vision and pattern recognition，1997

6 李国辉，胡峰，等. 基于内容的图像检索. 计算机世界，专题报道，1998

7 杨育彬，杭燕，王亮，伍静，陈世福：[基内容的信息智能检索系统技术报告]. 南京大学，2002


(上接第128页)

研究。

数据挖掘中抽样的应用和理论研究，将促进数据挖掘算法的改进，提高数据挖掘在分析大规模数据时的效率，最终推动数据挖掘的大规模实际应用。

## 参考文献


1 冯士雍，施锡铨. 抽样调查——理论、方法与实践. 上海科学技术出版社，1996

2 Kish L 著，倪加勋 译. 抽样调查. 中国统计出版社，1997

3 Olken F，Rotem D. Random sampling from databases：A survey Statistics and Computing，1995. 25～42

4 Toivonen H. Sampling large databases for association rules. VLDB'96，1996

5 Chen B，Haas P，Scheuermann P. New Two-Phase Sampling Based Algorithm for Discovering Association Rules，SIGKDD'02，2002

6 Megainduction J C. Machine Learning on Very Large Database：[PhD thesis]. School of Computer Science，University of Technology，Sydney，Australia，1991

7 Han Jiawei，Kamber M. 数据挖掘概念与技术. 机械工业出版社，2001

8 Palmer C R，Faloutsos C. Density biased sampling：An improved method for data mining and clustering. In：Proc. of the ACM SIGMOD'2000，2000

9 Kollios G，et al. Efficient Biased Sampling for Approximate Clustering and Qutlier Detection in Large Datasets；To Appear in IEEE TKDE Journal，2002

10 Provost F，Jensen D，Oates T. Efficient progressive sampling. In：Proc. of KDD'99，1999. 23～32

11 John G H，Langley P. Static versus dynamic sampling for data mining. In：Proc. of KDD'96，1996

12 Yang Y. Sampling strategies and learning efficiency in text categorization. In：Proc. of the AAAI Spring Symposium on Machine Learning in Information Access，1996. 88～95

13 Furnkranz J. Integrative windowing. Journal of Artificial Intelligence Research，1998. 129～164

14 Harris-Jones C，Haines T L. Sample size and misclassification：Is more always better；Working Paper of American Management Systems Center for Advanced Technologies，1997

15 Baohua GU，Feifang HU，Huan LIU. Sampling and Its Application in Data Mining：A Survey：[Technical report]. 2000