密度偏差抽样技术在聚类算法中的应用研究

余 波 朱东华 刘 嵩 郑 涛
(北京理工大学管理与经济学院 北京 100081)

摘 要 针对在大规模数据集上进行聚类困难的问题,分析了抽样技术的优点,研究了数据挖掘领域中的随机抽样的特点,并在此基础上提出了一种基于密度的偏差抽样方法。利用密度偏差抽样所获得的样本数据集能够较准确地反映总体数据集的特征,并且能够灵活地控制对数据集不同区域的抽样率。实验证明,在大规模数据集上进行聚类时,密度偏差抽样在时间复杂度上要优于随机抽样。

关键词 数据挖掘,聚类,偏差抽样,随机抽样
中图法分类号 TP311 **文献标识码** A

Applied Research on Clustering Algorithm Using Density Biased Sampling Technology

YU Bo ZHU Dong-hua LIU Song ZHENG Tao
(School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China)

Abstract The advantages of sampling technology were analyzed against the difficulties of clustering on large-scale data set, and study the traits of random sampling in data mining were studied then a biased sampling method based on density was presented. The sample data set using density biased sampling can more accurately reflect the character of the whole data set, and biased sampling can control the sampling rate freely as to different part of the data set. The experimental results show that, density biased sampling is superior to random sampling in time complexity when clustering on large-scale data set.

Keywords Data mining, Clustering, Biased sampling, Random sampling

在当前的数据挖掘领域中,大规模数据集越来越普遍,它们大多具有高维和海量的数据记录。针对这些日益增多的大型、高维数据集来说,现有数据挖掘方法的处理效果不太理想,并且处理大数据集系统的开销也比较大。由于大规模数据集的内在复杂性,因此在应用特定的数据挖掘方法,如聚类、关联规则等,对给定的数据集进行处理时,往往不是在整个数据集上进行处理,而是把抽样技术引入数据挖掘过程中,即先抽取出一个样本,然后在样本数据集上进行处理,最后根据处理结果来推测总体数据集的情况[1-3]。我们可以按照数据集的密度应用偏差抽样技术来加速聚类分析的运行,即在聚类分析过程中首先应用密度偏差抽样技术,依据数据分布的密度情况来生成样本数据集,然后进行聚类分析,在实现数据约简的情况下达到较好的聚类分析效果。

1 简单随机抽样在聚类挖掘中的应用

简单随机抽样(simple random sampling)是抽样方法的基础,在数据挖掘领域中应用广泛。Olken 等人对原始数据集需要进行简单随机抽样提供了一个极好的论据[4]。在专业的统计软件如 SAS,SPSS 中均包含简单随机抽样方法。

考查一个大小为 n 的数据集 D,其簇的大小也为 n,即所有 n 个点都是簇的成员。在这种情况下鉴别簇时,应用简单随机抽样就可以达到很好的效果。如果从簇中去除一些点并且引入相同数目的高斯噪声,尽管数据集的大小保持不变,但为了保证样本中同一区域的点像以前一样以相同的概率被包含在同一个簇中,就需要使用较大的样本。由此可知,包含在样本中的噪声数据给抽样技术带来了障碍。

Guha 等人提出了一个定理,该定理使样本大小 S 和簇中的一个区域被包含在样本中的概率结合在一起[5]。设 D 是一个大小为 n 的数据集,u 是一个大小为 $|u|$ 的簇,如果簇中有多于 $\Phi * |u|$ 个点在样本中,那么簇 u 被包含在样本中,其中 $0 \leqslant \Phi \leqslant 1$。

定理 1 对一个簇 u,如果样本大小 S 满足式(1),那么样本中属于簇 u 的点的个数小于 $\Phi * |u|$ 的概率小于 δ,$0 \leqslant \delta \leqslant 1$。式(1)可以由 Chernoff 界限来证明[6,7]。

$$S \geqslant \Phi n + \frac{n}{|u|}\log\left(\frac{1}{\delta}\right) + \frac{n}{|u|}\sqrt{\left(\log\left(\frac{1}{\delta}\right)\right)^2 + 2\Phi|u|\log\left(\frac{1}{\delta}\right)} \tag{1}$$

设某数据集有 $n=100000$ 个数据点,用式(1)来考查 S 对多个参数的灵敏度。图 1 和图 2 分别显示了 $\Phi=0.2$ 和 $\Phi=0.3$ 时 $|u|$ 与 S 的关系。

到稿日期:2008-03-20 本文受国家自然科学基金重点资助项目(70031010),985 哲学社会科学创新基地建设研究论文之一,"新世纪优秀人才支持计划"资助。

余 波(1978—),男,博士研究生,工程师,主要研究领域为数据挖掘、技术监测、模式识别、人工智能,E-mail:bitboyu@gmail.com;朱东华(1963—),男,博士生导师,研究员,主要研究领域为数据挖掘、技术监测、知识发现。

对比图1和图2可以看出，除非使用大样本，否则简单随机抽样不能以较高的概率保证一个小簇的一个大区域被包含在样本中。例如，为了用90%的可能性来保证有1000个点的簇的一个 $\Phi=0.2$ 的区域在样本中，我们需要对数据集进行20%的抽样；当在这个样本内的簇的区域增加到0.3时，所需的样本大小增长到将近数据集大小的31%。

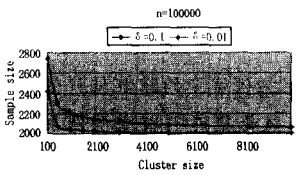

图1 $\Phi=0.2$ 时样本的变化

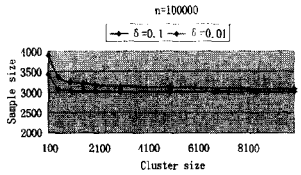

图2 $\Phi=0.3$ 时样本的变化

由以上分析可知，为了使简单随机抽样能够对聚类结果提供一个很好的保证，就必须使用大样本。这种要求就违背了抽样的原始动机，即约简数据总量。

2 偏差抽样在聚类挖掘中的应用

2.1 偏差抽样对比随机抽样

假设按照以下规则进行抽样：设 $\pi_1,\pi_2,\cdots,\pi_n$ 是数据集中的点，u 是一个簇，假定按照规则 R 使抽样过程存在偏差：

$$R:P(\pi_i \text{ 包含在样本中})=\begin{cases} p, & \pi_i\in u \\ 1-p, & \pi_i\notin u \end{cases}$$

其中 $0\leqslant p\leqslant 1$。

定理2 设 D 是一个大小为 n 的数据集，u 是大小为 $|u|$ 的簇，一个簇 u 被包含在样本中的条件是簇中有多于 $\Phi*|u|$ 个点被包含在样本中，其中 $0\leqslant\Phi\leqslant 1$。设 SR 是按照规则 R 进行抽样的样本，它可以保证 u 以超过 $1-\delta$ 的概率被包含在样本中，$0\leqslant\delta\leqslant 1$，$S$ 是用简单随机抽样得到的具有同样性能的样本。如果 $p\geqslant|u|/n$，那么

$$S_R\leqslant S \tag{2}$$

由式(1)和式(2)可以得到：

$$S_R>\frac{\Phi|u|+\log(\frac{1}{\delta})+\sqrt{(\log\frac{1}{\delta})^2+2\phi|u|\log(\frac{1}{\delta})}}{p} \tag{3}$$

对比式(1)和式(3)，可以得到如下结论：

定理2有一个非常直观的解释，即簇中的点以较高的概率被包含在样本中，因此可以更好地完成簇中的点有 Φ 部分被包含在样本中的要求，克服了简单随机抽样的局限。偏差抽样可以在相同概率的情况下使用较小的样本就能够把簇中相同的区域包含在样本中。

但是，没有明确的方法对簇中的点进行偏差抽样，因为不知道这些点的一个优先级。在此，引入概率密度函数。数据集的概率密度函数能够提供足够的信息来定位样本中的点，簇可以有效地被定位在稠密区域。通过抽样技术，按照数据集的密度，能够对稠密区域提高抽样概率以利于发现簇，因此还需要一个映射函数，在一定程度上把样本密度映射为抽样概率。

2.2 密度估算技术

一个密度估算函数试图定义一个近似数据分布的函数，它能够快速找到解决问题的近似方案。设 D 是一个有 d 个属性、n 个元组的数据集，设 $A=\{A_1,A_2,\cdots,A_d\}$ 是属性集，属性 A_i 服从正态分布。假定对一个属性序列，每一个元组是 d 维空间 $[0,1]^d$ 中的一个点。形式上，D 的一个密度估算函数是一个 d 维、非负函数：$f(x_1,\cdots,x_d)$，$f:[0,1]^d\to R$ 满足于：

$$\int_{[0,1]^d} f(x_1,\cdots,x_d)\mathrm{d}x_1,\cdots,\mathrm{d}x_d=1$$

理论上，偏差抽样技术可以使用任何密度估算方法，如对多维柱状图的计算[8]、应用核密度估算[9,10]等。由于核密度估算方法不利用有关数据分布的先验知识，对数据分布不附加任何假定，是一种从数据样本本身出发研究数据分布特征的方法。因此，使用核密度估算函数进行密度估算。核密度估算技术是基于统计方法尤其是核函数理论进行的[10]。为了讨论多维空间上的核密度估算问题，首先引入一维核估算的概念[11]。

定义1 设 $x_1,x_2,\cdots,x_n$ 为取值于 R 的独立同分布随机变量，它所服从的分布密度函数为 $f(x)$，$x\in R$，定义函数式(4)为密度函数 $f(x)$ 的核密度估算：

$$\hat{f}_h(x)=\frac{1}{nB}\sum_{i=1}^{n}K(\frac{x_i-x}{B}),x\in R \tag{4}$$

式中，$K(\cdot)$ 称为核函数，B 为预先给定的正数，通常称为带宽或光滑参数。

为方便起见，记 $Kh(u)=K(u/B)/B$，则式(4)可以表示为

$$\hat{f}_h(x)=\frac{1}{n}\sum_{i=1}^{n}K_h(x_i-x),x\in R \tag{5}$$

由定义1可知，分布密度函数 $f(x)$ 的核密度估算 $\hat{f}$ 不仅与给定的样本点集合有关，还与核函数的选择和带宽参数的选择有关，其中带宽参数 B 控制了在求点 B 处的近似密度时不同距离样本点对点密度的影响程度。由式(5)定义的核函数反映了分布密度函数在一点的值对该点领域样本点密度的正比依赖关系。为了密度函数估算的方便性与合理性，通常要求核函数满足以下两个条件：

(1) $K(-u)=K(u)$

(2) $Sup|K(u)|<\infty$，$\int_{-\infty}^{\infty}K(u)\mathrm{d}u=1$

常用的核函数有高斯核函数(Gaussian kernel)、Epanechnikov核函数和Biweight核函数。在聚类分析应用中，待处理的数据集一般是多维的。因此，需要讨论多维空间上的核密度估算问题。为方便起见，假设多维核函数是个同类型一维基核函数的乘积。虽然在文献[11]中证明了核函数的具体类型对结果的近似值影响不大，但为了处理问题的简便性，选择一个容易积分的核函数，即选择 Epanechnikov 核函数。计算一个核估算函数包括从数据集中选取一个大小为 S 的随机样本和计算数据集的每一维的标准差，这可以在一次数据扫描中完成。

2.3 密度偏差抽样技术

设 D 是一个有 n 个点的 d 维数据集，为简单起见，假设空间域为 $[0,1]^d$。设 $f(x_1,\cdots,x_d)$，$f:[0,1]^d\to R$ 是数据集 D 的一个密度估算函数。即对于一个给定区域 $R\subset[0,1]^d$，积分 $n\int_R f(x_1,\cdots,x_d)\mathrm{d}x_1\cdots\mathrm{d}x_d$ 近似等于区域 R 中的点的数量。定义点 x 的本地密度值为：

$$LD(x)=I/V$$

其中：$I=\int_B f(x)\mathrm{d}x$，B 是 x 邻域的一个半径为 $\varepsilon(\varepsilon>0)$ 的

球，V 是这个球的体积，$V=\int_B \mathrm{d}x$。

假设描述数据集密度的函数曲线是连续的且是光滑的，那么一个点周围的一个球内的本地密度代表了这个球内部的密度函数的平均值。因为函数曲线是光滑的，所以近似认为点 x 的本地密度值就是该点的函数值。假设数据集 D 的偏差样本具有如下性质：(1)D 中的一个点 x 被包含在样本中的概率是点 x 周围的数据空间的一个本地密度函数；(2)样本的期望大小是 b，b 是一个由用户设置的参数。

首先对简单随机抽样的实例进行考查。D 中的一个点被包含在样本中的概率是 b/n，根据密度估算函数的定义，如果 $f(x_1,\cdots,x_d)>1$，那么点 $(x_1,\cdots,x_d)$ 周围的本地密度大于该空间的平均密度，定义：

$$\frac{\int_{[0,1]^d} f(x_1,\cdots,x_d)\mathrm{d}x_1\cdots\mathrm{d}x_d}{V([0,1]^d)}=1$$

在密度偏差抽样中，我们想让一个给定点 $(x_1,\cdots,x_d)$ 包含在样本中的概率是该点周围空间的一个本地密度函数，它可由 $f(x_1,\cdots,x_d)$ 的值获得。另外，为了使偏差抽样技术具有可控制性，引入一个函数 $f^a(x_1,\cdots,x_d)=(f(x_1,\cdots,x_d))^a$，其中 a 为实数，该抽样技术的一个基本思想是抽样一个点 $(x_1,\cdots,x_d)$ 的概率与 $bf^a(x_1,\cdots,x_d)/n$ 对应成比例，因此 a 的值控制了偏差抽样的过程。

设 $k=\sum_{(x_1,\cdots,x_d)\in D} f^a(x_1,\cdots,x_d)$，定义 $f^*(x_1,\cdots,x_d)=\frac{n}{k}f^a(x_1,\cdots,x_d)$，则 f^* 具有如下附加特性：

$$\sum_{(x_1,\cdots,x_d)\in D} f^*(x_1,\cdots,x_d)=n$$

在抽样过程中，每一点 $(x_1,\cdots,x_d)\in D$ 被包含在样本中的概率是 $bf^*(x_1,\cdots,x_d)/n$，即

$$b(f(x_1,\cdots,x_d))^a/k$$

很明显地可以看出，该算法满足性质 1。因为期望的样本大小为

$$\sum_{(x_1,\cdots,x_d)\in D}\frac{b}{n}f^*(x_1,\cdots,x_d)$$

并且

$$\sum_{(x_1,\cdots,x_d)\in D}\frac{b}{n}f^*(x_1,\cdots,x_d)=\frac{b}{n}\sum_{(x_1,\cdots,x_d)\in D}f^*(x_1,\cdots,x_d)=b$$

所以该算法也满足性质 2。对该算法的描述如图 3 所示。

> 给定一个数据集 D，一个数据集密度估算函数 f 和一个用户设置的样本大小 b。
> 1) 计算被包含在样本中的每一点 x 的密度偏差概率。
> 2) 对数据集进行一次扫描，按相应概率输出每一点。

图 3　偏差抽样技术

由以上可知，该算法需要对数据集进行两次扫描。假设密度估算函数 f 可用，k 的值在第一次扫描数据集时可以被计算出来。在第二次扫描过程中，对数据集中的每一点 x，能够计算出 $f^*(x)$ 的值。该技术的一个重要的优点是它非常灵活，可以适用于不同的应用环境。随着 a 值的变化，既能对高密度区域进行充分抽样（$a>0$ 时），也能对低密度区域进行充分抽样（$a<0$ 时）。具体描述如下：

$a=0$ 时，每个点被包含在样本中的概率为 b/n。在这种情况下，得到的是简单随机抽样，由此可见，简单随机抽样是偏差抽样的一种特殊情况。

$a>0$ 时，设 x 和 y 分别是位于数据集中不同密度区域的点集，如果 $f^a(x)>f^a(y)$，那么 $f(x)>f(y)$，表示在高密度区域的抽样率要比低密度区域的抽样率高。准确来说，在数据空间中，如果某区域的密度高于平均密度，那么偏差抽样在该区域的抽样率比简单随机抽样的抽样率要高；同理，如果该区域的密度低于平均密度，那么偏差抽样在该区域的抽样率比简单随机抽样的抽样率要低。这说明对数据密集区的抽样可以使它在样本中依然密集。

$a<0$ 时，与 $a>0$ 中的情况相反。如果 $f^a(x)>f^a(y)$，那么 $f(x)<f(y)$，表示低密度区域的抽样比简单随机抽样高，而高密度区域的抽样比简单随机抽样低。这意味着，用这样的样本更有可能发现小的和稀疏的簇。

3　时间复杂度分析

在有 n 个数据点的数据集上分别进行随机抽样和偏差抽样，在样本数据集上运行 CURE 聚类算法，对随机抽样与偏差抽样的时间复杂度进行分析。由于 CURE 的运行时间是二次的，构造密度函数进行的一次扫描数据集所花费的时间可以通过在一个较小的样本上运行一个二次算法来弥补。例如，如果在一个 1% 的随机样本上运行 CURE 聚类算法，$O(n)$ 是选取随机样本所需的时间，$O(n/100)^2$ 是在随机样本上执行 CURE 所需的时间，则在 1% 的随机样本上运行 CURE 聚类算法总的运行时间是 $O(n+(n/100)^2)$。如果在一个 0.5% 的偏差样本上运行CURE 聚类算法，$O(2n)$ 是选取偏差样本所需的时间，$O(n/200)^2$ 是在偏差样本上执行 CURE 所需的时间，则在 0.5% 的偏差样本上运行 CURE 聚类算法总的运行时间是 $O(2n+(n/200)^2)$。表 1 显示了在 n 不同时的时间复杂度。

表 1　时间复杂度

n	1000	5000	10000	13334	30000	50000	100000
随机样本	1100	7500	20000	31113	120000	300000	1100000
偏差样本	2025	10625	22500	31113	82500	162500	450000

由表 1 可以看出，当 $n>13334$ 时，偏差抽样表现出比随机抽样优越的性能，并且随着 n 的变大，这种优越性变得更加明显。这就从定量的角度证明了在大规模数据集上进行聚类时，应用偏差抽样得到的偏差样本比随机样本更能满足用户的要求。

结束语　通过对简单随机抽样在聚类挖掘过程中存在缺点的分析，提出了一种基于密度的偏差抽样技术。该技术在非对称、不均匀分布的数据集上进行抽样时，所获得的样本数据集能够较准确地反映总体数据集的特征，并且能够灵活地控制对数据集不同区域的抽样率。在偏差样本上进行聚类分析，可以更准确地找出原始数据集中的簇，从而有利于决策者进行相关的分析并做出正确的决策。

参 考 文 献

[1] Toivonen H. Sampling large databases from association rules// VLDB'96. 1996

[2] Chen B, Haas P, Scheuermann P. New Two-phase Sampling-based Algorithm for Discovering Association Rules// SIGKDD'02. 2002

（下转第 264 页）

把图像缩小为 16×16，并把它变换为一列，这样就得到 256×40 的 Gabor 特征图片，在此基础上进行识别。这样在 Georgia Tech 库上的识别率比较如图 8 所示，而对应的主元数筛选情况如图 9 所示。

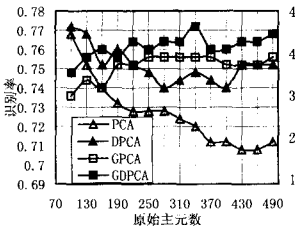

图 8 Georgia Tech 库上的识别率比较图

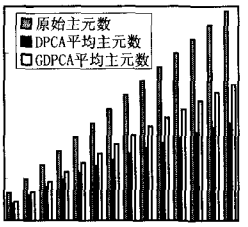

图 9 Georgia Tech 库上的主元数比较图，其中原始主元数对应于图中的原始主元数

从实验结果可以看出，利用本文提出的 DPCA 和 GDPCA 算法，与相应文献[2]中用的 PCA 算法和文献[11]中用的 GPCA 算法相比，在主元有一定数量减少的情况下，识别率却有所提高。从图 6 和表 1 可以看出，在 ORL 人脸库上 DPCA 和 GDPCA 算法最高识别率分别为 94.5%和 97.5%，平均识别率也均比 PCA 和 GPCA 算法高。从图 8 和表 1 可以看出，在 Georgia Tech 人脸库上由于样本自身姿态的变化比较大，所以最高识别率不高，但我们的 DPCA 和 GDPCA 算法仍然比相应 PCA 和 GPCA 的算法在平均识别率还是在最高识别率上均有相应的提高。从图 7 和图 9 可以看出，在两个人脸库上，在识别率提高的情况下，本文提出的算法所用的主元数均少于传统的算法。这说明本文提出的动态主成分子空间构造算法，在选择和优化特征空间，去除干扰特征方面具有很好的效果。

表 1 平均识别率对照表

算法	R_1(%)	R_2(%)
PCA[2]	93.06	72.66
DPCA	93.65	75.17
GPCA[11]	95.76	75.11
GDPCA	96.53	76.14

注：R1 代表 ORL 库上的平均识别率，R2 代表 Georgia Tech 库上的平均识别率。

结束语 本文根据待测试人脸图像的个体差异性，利用多元线性回归分析的数学统计理论，对 PCA 算法初步选定的特征空间进一步筛选，提出了一种自主式动态的特征选择方法。为了验证效果，我们首先把该算法用到传统的 PCA 算法中，提出了 DPCA 算法，在 ORL(AT&T) 和 Georgia Tech 人脸数据库的实验表明：本文的动态主成分子空间构造算法，在主元数目减少的情况下，而识别率却有进一步的提高；其次，为了验证算法的普实性，把该算法应用到基于 GPCA 中，提出了 GDPCA 算法，在 ORL(AT&T)和 Georgia Tech 人脸数据库的实验中同样取得了很好的效果。这些实验结果充分说明了本算法对特征选择和优化的有效性。

参 考 文 献

[1] Kirby M, Sirovich L. Application of the Karhunen-Loéve procedure for the characterizatio-n of human faces. IEEE Trans. Pattern Analysis and Machine Intelligence, 1990, 12(1): 103-108

[2] Turk M, Pentland A. Eigenfaces for Recogni - tion. Journal of Cognitive Neuroscience, 1991, 3(1): 71-86

[3] Belhumeur P N, Hespanda J, Kriegeman D. Eigenfaces vs. Fiserfaces: Recognition Using Class Special Linear Projection. IEEE Trans. Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720

[4] Moghaddam B, Jebara T, Pentland A. Bayesian face recognition. Pattern Recognition, 2000, 13(11): 1771-1782

[5] Wang Xiaogang, Tang Xiaoou. A unified fra - mework for subspace face recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 2004, 26 (9): 1222-1228

[6] Bartlett M S, Movellan J R, Sejnowski T J. Face Recognition by Independent Component Analysis. IEEE Trans. on Neural Networks, 2002, 13(6): 1450-1464

[7] Gong Xun, Wang Guoyin. A Dynamic Comp - onent Deforming Model for Face Shape Rec-onstruction// Proceeding of the International Symposium on Visual Computing 2007, L-NCS 4841. US, 2007: 488-497

[8] Chen Songcan, Zhang Daoqiang, Zhou Zhihua. Face recognition with one training image per person. Pattern Recognition Letters, 2002, 23(14): 1711-1719

[9] 王蕴红，范伟，谭铁牛. 融合局部与全局特征的子空间人脸识别. 计算机学报，2005，28(10)：1657-1663

[10] Jones J P, Palmer L A. An evaluation ofthe two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of Neurophysiology, 1987, 58(6): 1233-1258

[11] Liu Chengjun, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. on Image Processing, 2002, 11(4): 467-476

[12] Zhang Baochang, Shan Shiguang, Chen Xilin, et al. Histogram of Gabor Phase Patterns(HGPP): A Novel Object Representation Approach for Face Recognition. IEEE Trans. on Image Processing, 2007, 16(1): 57-68

(上接第 209 页)

[3] 张春阳，蔡庆生，等. 抽样在数据挖掘中的应用研究. 计算机科学，2004，31(2)

[4] Olken F, Rotem D, Xu Ping. Random sampling from hash files// Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data. ACM Press, 1990: 375-386

[5] Guha S, Rastogi R, Shim K. CURE: An Efficient Clustering Algorithm for Large Databases// Proc. ACM SIGMOD Conf.. June 1998: 73-84

[6] Knorr E, Ng R. A unified notion of outliers: Properties and computation// Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining(KDD'97). Newport Beach, CA. Aug. 1997: 219-222

[7] Motwani R, Raghavan P. Randomizeed Algorithms. Cambridge University Press, 1995

[8] Poosala V, Ioannidis Y. Selectivity Estimation Without the Attribute Value Independence Assumption// Proc. Very Large Data Bases Conf.. Aug. 1997: 486-495

[9] Blohsfeld B, Korus D, Seeger B. A Comparison of Selectivity Estimators for Range Queries on Metric Attributes// Proc. ACM SIGMOD Int'l Conf. Management of Data. 1999

[10] Scott D. Multivariate Density Estimation: Theory, Practice and Visualization. Wiley& Sons, 1992

[11] 李存华，孙志挥. 核密度估算及其在聚类算法构造中的应用. 计算机研究与发展，2004，41(10)：1713-1719