DBSCAN算法中参数的自适应确定

李宗林,罗 可
LI Zonglin, LUO Ke

长沙理工大学 计算机与通信工程学院,长沙 410114
Institute of Computer and Communication Engineering, Changsha University of Sciences and Technology, Changsha 410114, China

LI Zonglin, LUO Ke. Research on adaptive parameters determination in DBSCAN algorithm. Computer Engineering and Applications, 2016, 52(3):70-73.

Abstract: DBSCAN algorithm needs *Eps* and min*Pts* two parameters, leading to the accuracy of clustering results directly depends on the user's choice of parameters, thus this paper puts forward a new method of parameter determination. It adopts nonparametric kernel density estimation theory to analyse the distribution features of the data samples to automatically determine the *Eps* and min*Pts* parameters, avoiding the manual intervention of clustering process, and achieving automation of clustering process. Theoretical analysis and experimental results show that this method is able to choose reasonable parameters of *Eps* and min*Pts* and clustering results with higher accuracy are obtained.

Key words: Density Based Spatial Clustering of Applications with Noise(DBSCAN); kernel density estimation; self-adaptive; clustering

摘 要: DBSCAN算法需要人为确定 *Eps* 和 min*Pts* 两个参数,导致聚类结果的准确度直接取决于用户对参数的选择,因此提出一种新的参数确定方法,采用非参数核密度估计理论分析数据样本的分布特征来自动确定 *Eps* 和 min*Pts* 参数,避免了聚类过程的人工干预,实现聚类过程的自动化。理论分析和实验结果表明,该方法能够选择合理的 *Eps* 和 min*Pts* 参数,并得到了较高准确度的聚类结果。

关键词: 一种经典的基于密度的聚类算法(DBSCAN);核密度估计;自适应;聚类
文献标志码: A **中图分类号:** TP301 doi:10.3778/j.issn.1002-8331.1402-0278

1 引言

DBSCAN算法是一种经典的基于密度的聚类算法[1],它以单位超球状区域内所包含数据对象的数量来衡量此区域密度的高低。DBSCAN算法能够发现任意形状的簇,并有效识别离群点,但聚类之前需要人工选择 *Eps* 和 min*Pts* 两个参数。已有一些文献提出了若干判定的方法。文献[2]中给出的参数选择方法是设定 min*Pts*=4,通过观察法来判断 *Eps*。显然,这种方法需要用户的参与。文献[3]的研究中引入了距离分布的概念。仍假定 min*Pts*=4,计算数据集中两个对象的距离并排序,取 *Eps* 为排序后第 $\delta^2/(4c)C_n^2$ 个点对应的值,这里仍需设定簇个数 *c* 的值。文献[4]通过增加簇连接信息使DBSCAN对输入参数的敏感性降低,但未能解决自动确定参数的问题。文献[5]提出了逐级细化聚类的方法,每次聚类动态调整参数,但初始参数仍需给定。文献[6]根据 *K*-dist图的思想,计算每个数据第 *k* 个最近邻的距离并排序,虽然可以确定分区的 *Eps* 值,但仍需指定 min*Pts*。文献[7]提出的I-DBSCAN算法通过分析数据集的统计特性来确定两个参数的方法,虽然达到了自动聚类的目的,但依赖数据的本身属性,而且通过观

基金项目: 国家自然科学基金(No.11171095,No.71371065);湖南省自然科学衡阳联合基金(No.10JJ8008);湖南省科技计划项目(No.2013SK3146)。

作者简介: 李宗林(1988—),男,硕士,主要研究方向为数据挖掘,E-mail:lizonglin10086@163.com;罗可(1961—),男,博士,教授,主要研究方向为数据挖掘、计算机应用等。

收稿日期: 2014-02-26 **修回日期:** 2014-04-29 **文章编号:** 1002-8331(2016)03-0070-04
CNKI网络优先出版:2014-06-24, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1402-0278.html

察分析去拟合理论上存在误差。

鉴于此,本文提出了一种新的参数确定方法,采用非参数核密度估计理论分析数据样本的分布特征来自动确定 Eps 和 $\min Pts$ 参数,由于非参数密度估计能够处理任意的密度分布,无需对样本分布的形式做出假设,仅以数据点作为概率密度函数估计的依据,因此得到的聚类结果更加准确。

2　预备知识

2.1　DBSCAN算法

DBSCAN是一种经典的基于密度的聚类算法,可以自动确定簇的数量,并能够发现任意形状的簇。DBSCAN主要定义如下:

定义1(Eps 邻域) 给定一个数据对象 p,p 的邻域 $N_{Eps}(p)$ 定义为以 p 为核心,以 Eps 为半径的 d 维超球体区域,即

$$N_{Eps}(p)=\{q\in D|dist(p,q)\leqslant Eps\} \tag{1}$$

其中, $D\in R^d$ 为 d 维实空间上的数据集, $dist(p,q)$ 表示 D 中的2个对象 p 和 q 之间的距离。

定义2(核心点与边界点) 对于对象 $p\in D$,给定一整数 $\min Pts$,若 $|N_{Eps}(p)|\geqslant \min Pts$,则称 p 为核心点;非核心点但在某核心点的 Eps 邻域内的对象称为边界点。

定义3(直接密度可达) 给定 $(Eps,\min Pts)$,若满足: $p\in N_{Eps}(q)$; $N_{Eps}(q)\geqslant \min Pts$,则称对象 p 是从 q 出发,直接密度可达的。

定义4(密度可达) 给定数据集 D,当存在对象 $p_1,p_2,\cdots,p_n$,其中, $p_1=q$, $p_n=p$,对于 $p_i\in D$,若在 $(Eps,\min Pts)$ 下 p_{i+1} 从 p_i 直接密度可达,则称对象 p 从 q 密度可达。密度可达是非对称的。

定义5(密度相连) 若数据集 D 中存在对象 o,使得对象 p 和 q 是从 o 密度可达的,那么称对象 p 和 q 密度相连。密度相连是对称的。

定义6(簇和噪声) 由任意一个核心点对象开始,从该对象密度可达的所有对象构成一个簇,不属于任何簇的对象为噪声。

2.2　核密度估计

核密度估计(kernel density estimation)是在概率论中用来估计未知的密度函数,属于非参数检验方法之一,由Rosenblatt和Parzen提出,又名Parzen窗(Parzen window)[8]。由于核密度估计方法不利用有关的数据分布的先验的知识,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。因而在统计学理论和应用领域受到高度重视。该方法把每个观察对象看作一个周围区域中的高概率密度指示器,一个点上的概率密度依赖于该点到观察对象的距离[9]。

为了讨论多维空间上的核密度问题,首先引用一维核估计的概念[10]。

定义7　设 $x_1,x_2,\cdots,x_n$ 为取值于 R 的独立同分布随机变量,其所服从的分布密度函数为 $f(x)$, $x\in R$。定义函数:

$$\hat{f}_h(x)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{2}$$

$\hat{f}_h(x)$ 称为密度函数 $f(x)$ 的核密度估计, $K\left(\frac{x-x_i}{h}\right)$ 为核函数, h 称为窗宽或光滑参数, n 为样本数量。

定义8　设 $x_1,x_2,\cdots,x_n$ 为空间 R^d (d维)上的独立同分布随机变量,其分布密度函数 $f(x)$ 的核密度估计定义为:

$$\hat{f}_h(x)=\frac{1}{n}\sum_{i=1}^{n}K_{d,h}(x-x_i) \tag{3}$$

其中,

$$K_{d,h}(x)=\prod_{i=1}^{d}K(x_i/h_i)/h_i \tag{4}$$

3　参数的自适应确定

在DBSCAN算法中,密度通过统计被参数 Eps 定义的邻域中的对象个数来计算。这种密度估计对半径值非常敏感,随着半径的稍微增加,密度显著改变。为了解决这一问题,可以使用核密度估计。为方便起见,记 $K_h(u)=K(u/h)/h$,则式(2)可以表示为:

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

在理论上,任何函数均可以用做核函数,但为了密度函数的方便性和合理性,通常要求核函数满足以下条件:

$$K(-u)=K(u) \tag{6}$$

$$\int_{-\infty}^{\infty}K(u)\mathrm{d}u=1 \tag{7}$$

经常使用的核是均值为0,方差为1的标准高斯函数:

$$K(u)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{\frac{u^2}{2}} \tag{8}$$

特别地,采用 d 维同参数高斯核函数构造的 R^d 上的核密度函数具有如下形式:

$$K_{d,\sigma}(x)=(\sqrt{2\pi}\sigma)^{-d}\mathrm{e}^{-\frac{\sum((Y_i-x_i)^2)}{2\sigma^2}} \tag{9}$$

3.1　Eps 的自适应确定

在核密度估计理论中,带宽值的选择对估计量 $\hat{f}_h(x)$ 的影响很大,如果 h 太小,那么密度估计局限于观测数据的附近区域,致使估计密度函数出现很多错误的峰

值,在聚类分析应用中表现为一个自然簇被错误地拆分成多个簇。如果 h 太大,那么密度估计就把概率密度分开得太散,这样会过滤掉一些重要特征,在聚类分析应用中表现为噪声被错误地归入簇,若干个自然簇也被错误地合并成一个簇。

依据上面的分析,假如能够通过数学方法确定 h 的大小,则可以确定参数 Eps 的值。即

$$Eps=h \tag{10}$$

统计学上,通常使用积分均方误差 $MISE(h)$ 作为判断密度估计量好坏的准则。

$$MISE(h)=AMISE(h)+O\left(\frac{1}{nh}+h^4\right) \tag{11}$$

其中,

$$AMISE(h)=\frac{\int K^2(x)\mathrm{d}x}{nh}+\frac{h^4\sigma^4\int[f''(x)]^2\mathrm{d}x}{4} \tag{12}$$

要最小化 $AMISE(h)$,必须把 h 设在某个中间值,才可以避免 $\hat{f}_h(x)$ 有过大的偏差(太过光滑)。关于 h 最小化 $AMISE(h)$,最好是精确地平衡 $AMISE(h)$ 中偏差项和方差项的阶数。

最优的窗宽是[11]:

$$h=\left(\frac{\int K^2(x)\mathrm{d}x}{n\sigma^4\int[f''(x)]^2\mathrm{d}x}\right)^{1/5} \tag{13}$$

简便起见,定义:

$$R(g)=\int g^2(z)\mathrm{d}z \tag{14}$$

针对最小化 $AMISE(h)$ 得到的最优带宽中含有未知量 $R(f'')$,Sliverman提出了拇指法则(rule of thumb)[10]:把 f 用方差和估计方差相匹配的正态密度替换。这就等于用 $R(\phi'')/\hat{\sigma}$ 估计 $R(f'')$,其中 ϕ 为标准正态密度函数。若取核函数为高斯密度核函数,σ 使用样本方差 $\hat{\sigma}$,利用拇指法则得到:

$$h=(4/3n)^{1/5}\hat{\sigma} \tag{15}$$

3.2　min Pts 的自适应确定

根据核密度估计理论,假设 R 是一个以某对象为中心,边长为 l 的极小立方体(或者半径为 l 的极小球体),现在要考虑的是落入立方体数据点的个数 N。

定义一个核函数:

$$K(u)=\begin{cases}1, |u_i|\leqslant\frac{1}{2}\\ 0, |u_i|>\frac{1}{2}\end{cases} \tag{16}$$

其中 $i=1,2,\cdots,d$。

该函数的意义是:数据维数为 d 维,当样本数据点落入极小立方体时,函数值为1;其他情况下为0。所以落入立方体数据点的总个数 N 就可以表示为:

$$N=\sum_{i=1}^{n}K\left(\frac{x-x_i}{l}\right) \tag{17}$$

上式的思想运用在DBSCAN算法中可以这样理解:以对象 x 为中心、以 h 为半径的空间内存在的对象个数为 min Pts,当用 h 替换 l,就得到 min Pts 的值。即

$$\min Pts=\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{18}$$

4　实验及结果比较

4.1　聚类结果

实验环境:操作系统 Windows7,软件:Microsoft Visual Studio 2010,Matlab 2012,硬件:Inter Core i3 CPU,内存4 GB。为了验证本文算法的有效性和可行性,使用本文算法对数据集SampleD、DS1和DS2分别进行聚类。在进行算法有效性验证时,本文采用有监督的F度量(F-Measure)方法[12]来检测。SampleD是一个240个对象的二维数据集,DS1是一个520个对象的二维数据集,DS2是一个300个对象的二维数据集。三者的聚类结果分别如图1~3所示。由图1~图3可以看到,自适应算法能够发现数据集中的高密度区域并做出适当的簇划分。这表明本文算法能够有效选择合适的 Eps 和 min Pts 参数。

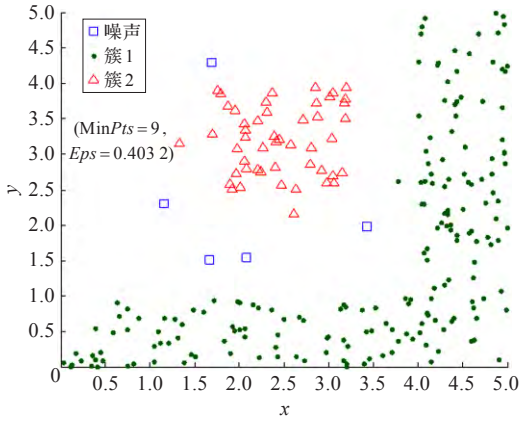

图1　SampleD的聚类结果

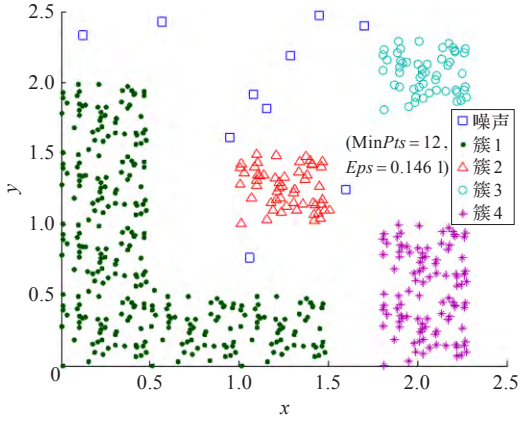

图2　DS1的聚类结果

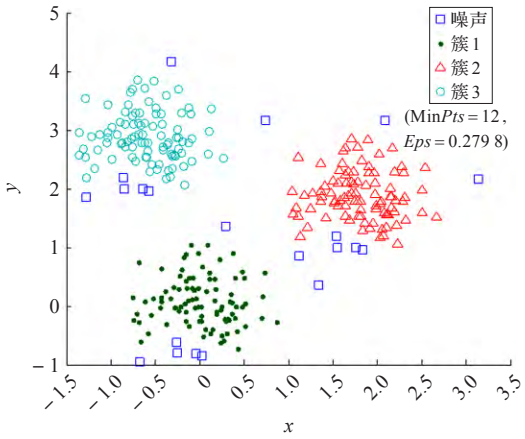

图3　DS2的聚类结果

SampleD、DS1和DS2数据集的各项聚类结果和准确率在表1给出。为了测试本文算法在二维以上数据集的应用效果，使用Iris数据集进行实验，因多维空间绘图不便，只给出聚类结果的数据形式，见表1。为了比较，对以上数据集同时进行I-DBSCAN算法和传统DBSCAN算法聚类。其中，DBSCAN取 $\min Pts=4$，$Eps=Eps_4$（Eps_4 指 $\min Pts=4$ 时的 Eps 值）。从表1可以看出，直接取 $\min Pts=4$，$Eps=Eps_4$ 进行DBSCAN聚类的准确率不高，因此对参数进行判断而不是取固定值是必要的。

表1　聚类算法的准确度

数据集	维数	对象数	聚类算法	Eps	$\min Pts$	准确度/%
SampleD	2	240	本文算法	0.403 2	9	93.36
			I-DBSCAN	0.468 1	10	88.23
			DBSCAN	0.234 7	4	45.41
DS1	2	520	本文算法	0.146 1	12	97.86
			I-DBSCAN	0.143 5	11	90.47
			DBSCAN	0.071 1	4	41.68
DS2	2	300	本文算法	0.279 8	12	97.23
			I-DBSCAN	0.291 2	12	91.85
			DBSCAN	0.154 8	4	63.95
Iris	4	150	本文算法	0.530 4	8	93.68
			I-DBSCAN	0.400 3	7	88.03
			DBSCAN	0.319 4	4	69.50

从准确率来看，无论是在处理包含超球状簇的数据集（DS2和Iris）还是任意形状簇的数据集（SampleD和DS1）的时候，本文算法和I-DBSCAN算法都有着不错的表现，本文算法的准确率更高，这得益于DBSCAN算法本身具有的优势。

对于较高维数据集（Iris），传统的DBSCAN算法和I-DBSCAN算法都显得效果不理想，这是因为高维数据集中数据对象的分布更加随机，对象之间欧几里德距离的差异变得不明显，可能导致整个数据集被聚成单一的一个簇。而本文算法采用非参数密度估计理论，由于其能够处理任意的密度分布，无需对样本分布的形式做出假设，因此得到的聚类结果更加准确。

4.2　进一步讨论

对于密度差别很大的数据集[13]，三种算法的聚类效果都不怎么理想。这是基于密度的聚类算法本身存在的问题。传统的DBSCAN算法使用全局单一的参数 Eps 和 $\min Pts$，导致聚类过程中只有一个密度衡量指标，如果 Eps 选择过大，高密度的自然簇可能被合并；选择过小则低密度的自然簇被丢弃。

I-DBSCAN算法虽然不会出现低密度自然簇被大面积丢弃的现象，但有可能造成高密度自然簇的合并。本文算法虽然不会出现上述的极端情况，但核密度估计在估计边界区域的时候有可能出现边界效应。

本文算法利用核密度估计一组样本，$\hat{f}_h(x)$ 的计算量随着样本数量的增大而增大，虽然增加了算法的准确率，却也牺牲了时间复杂度。对于多数实际问题，可以考虑将整个样本数据划分成多个等距网格小区间[14]。虽然准确率会稍有影响，却也在准确率和复杂度之间取得一个平衡。

5　结束语

DBSCAN是一种经典的基于密度的聚类算法，可以自动确定簇的数量，并能够发现任意形状的簇。由于DBSCAN算法需要人工输入 Eps 和 $\min Pts$ 两个参数，导致聚类准确率较低。本文在DBSCAN的基础上，提出了自适应确定 Eps 和 $\min Pts$ 参数的方法，通过核密度估计理论建立合适的数学模型判断 Eps 和 $\min Pts$，此过程无需人工输入参数，能够实现聚类过程的全自动化。由于无需任何有关当前数据集的先验知识，全靠数学模型驱动，所以自动聚类的准确率较高，在各个应用领域能够发挥重要作用，特别是对于在线数据的实时聚类有重要意义。当然，本文算法也存在一些不足：对于密度差别很大的数据集，聚类效果一般；算法计算复杂度为 $O(n^2)$，对于处理大数据，成本太高。因此，如何有效解决这些问题将是下一步的研究方向。

参考文献：

[1] Han Jiawei，Kamber M.数据挖掘：概念与技术[M].范明，译.北京：机械工业出版社，2012：306-309.

[2] Ester M，Kriegel H P，Sander J.A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Simoudis E.Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. Portland：AAAI Press，1996：226-231.

[3] Yue S H，Li P，Guo J D，et al.A statistical information-based clustering approach in distance space[J].Journal of Zhejiang University：Science，2005，6(1)：71-78.

（下转80页）

其真实值,算法的整体误差也就越小。然而,由于算法是在检测到收敛后才对之前的观测信号进行回复分离,若阈值取值过小,相应的分离信号与观测信号之间就会存在很大的时延差。所以,实际应用时,应根据所需的分离精度与所允许的时延两方面对算法中的阈值进行折中选取。

由图11中可以看出,当阈值 α_{opt} 取值足够较小时,改进的NA-EASI算法分离出的信号波形能够很好地逼近于源信号。该算法能够准确检测到信道的变化并在线估计信源的个数,其性能未受信道变化或是信源个数变化的影响,在整个分离过程中都保持着较高的分离精度。只是在信道变化或是信源个数变化后,分离信号的排列次序发生了变化,然而这并不影响对源信号波形信息的获取。

6 结束语

本文利用EASI算法中估计函数的期望作为在线调整步长的依据(步长控制因子),提出了一种新的步长自适应算法——NA-EASI算法,其与传统EASI算法相比,在收敛速率和稳态误差方面均有较大改善。

为了能够进一步提高信号的分离精度,解决时变系统中不同条件下的盲源分离问题,本文建立了一种混合矩阵变化的在线检测机制,并将这种在线检测机制与NA-EASI算法相结合,提出了一种改进的NA-EASI算法。仿真实验证明,改进的NA-EASI算法能够提高分离初期或是信道变化后分离初期的信号分离精度,解决源信号为非零均值信号时的盲源分离问题,并且能够准确地在线估计源信号的个数,实现信源数变化条件下的盲源分离。

参考文献:

[1] 孙守宇.盲信号处理基础及其应用[M].北京:国防工业出版社,2010.

[2] Cardoso J F,Laheld B H.Equivariant adaptive source separation[J].IEEE Trans on Singal Processing,1996,44(12):3017-3029.

[3] Yang H H.Serial updating rule for blind separation derived from the method of scoring[J].IEEE Trans on Signal Processing,1999,47(8):2279-2285.

[4] Yuan L X,Wang W W.Variable step-size sign natural gradient algorithm for sequential blind source separation[J].IEEE Signal Processing,2005,12(8):589-592.

[5] 朱孝龙.盲自适应信号分离的并行实现方法研究[D].西安:西安电子科技大学,2002:1-35.

[6] 李广彪,张建云.基于分离度的步长自适应自然梯度算法[J].信号处理,2007,23(3):429-432.

[7] 付卫红,史凡,刘乃安.适用于时变信道环境的盲源分离算法[J].电子科技大学学报,2012,41(4):512-515.

[8] Ou Shifeng,Gao Ying,Jin Gang,et al.Variable step size algorithm for blind source separation using a combination of two adaptive separation systems[C]//International Conference on Natural Computation,2009:649-652.

[9] 王荣杰,周海峰,詹宜巨,等.一种基于牛顿迭代的自适应复盲源分离算法[J].电子学报,2014,42(6):1125-1131.

[10] 陈海平,张杭,张江.回溯式在线EASI盲源分离算法[J].信号处理,2013,29(9):1250-1255.

[11] 蒋照菁,辜方林,张杭.一种基于NPCA的自适应变步长盲源分离算法[J].计算机工程与应用,2013,49(8):206-208.

[12] von Hoff T P,Lindgren A G,Kaelin A N.Step-size control in blind source separation[C]//Independent Component Analysis and Blind Signal Separation,2000:509-514.

[13] 冶继民,张贤达,朱孝龙.信源个数未知和动态变化时的盲信号分离[J].中国科学:E辑,2005,35(12):1277-1287.

[14] Amari S,Cichocki A,Yang H H.A new learning algorithm for blind signal separation[J].Neural Information Processing Systems,1996,8:757-763.

(上接73页)

[4] 蔡颖琨,谢昆青,马修军.屏蔽了输入参数敏感性的DBSCAN改进算法[J].北京大学学报:自然科学版,2004,40(3):480-486.

[5] 苏中,马少平,杨强,等.基于Web-Log Mining的Web文档聚类[J].软件学报,2002,13(1):99-104.

[6] 余亚飞,周爱武.一种改进的DBSCAN密度算法[J].计算机技术与发展,2011,21(2):30-33.

[7] 周红芳,王鹏.DBSCAN算法中参数自适应确定方法的研究[J].西安理工大学学报,2012,28(3):289-292.

[8] 王星.非参数统计[M].北京:清华大学出版社,2009.

[9] 李大威,徐立宏.一种迭代的核密度估计视觉目标检测算法[J].系统仿真学报,2013,25(3):558-564.

[10] Hall P,Wand M.On the accuracy of binned kernel density estimators[J].Journal of Multivariate Analysis,1994,56(2):165-184.

[11] 黎运发,黄名辉.核密度估计逐点最优窗宽选择的改进[J].统计与决策,2011,14(7):28-32.

[12] Steinbach M,Karypis G,Kumar V.A comparison of document clustering techniques[C]//KDD Workshop on Text Mining,2000:525-526.

[13] 周董,刘鹏.VDBSCAN:变密度聚类算法[J].计算机工程与应用,2009,45(11):137-141.

[14] 董琰,葛君伟.一种基于网格密度的自适应聚类分析算法[J].计算机应用研究,2007,24(8):56-66.