DOI:10.13718/j.cnki.xsxb.2019.05.017

基于密度聚类的数据库入侵检测系统研究①

曹德胜

华北科技学院 计算机学院,北京 065201

摘要:针对现有数据库入侵检测系统高误报率的问题,提出了一种基于密度聚类数据库入侵检测系统,其检测系统过程分为 2 个部分,①数据训练阶段:执行事务属性的数据预处理,然后将数据集划分为训练集和测试集,使用点排序识别聚类结构(Ordering of Points To Identify Clustering Structure, OPTICS)来构建用户的正常配置文件;②入侵检测阶段:每个传入行为有 2 种状态,位于群集内或是集群外,根据其局部异常因子(Local Outlier Factor, *LOF*)值来确定事务的异常程度,对于 $LOF<1$ 的行为允许访问数据库,其他行为通过采用不同的监督机器学习技术进一步验证是正常值或异常值,实现入侵检测. 实验结果表明,与其他现有数据库入侵检测系统相比,本文系统性能优于其他 2 种系统.

关 键 词:入侵检测;密度聚类;点排序识别聚类结构;局部异常因子;监督学习

中图分类号:TP392 **文献标志码**:A **文章编号**:1000-5471(2019)05-0103-06

随着大数据时代的到来[1-2],收集和存储在数据库中的数据量也以惊人的速度快速增长,随之增长的是入侵活动和安全攻击[3-5]. 标准数据库安全机制以及基于网络和基于主机的入侵检测系统已经无法检测专门针对数据库的恶意攻击. 数据库系统中的入侵攻击可分为外部攻击和内部攻击,外部人为获取数据库而进行的恶意交易称为外部攻击,组织内用户发生的攻击意识到安全性设置并具有某些资源的访问权限称为内部攻击[6].

每个用户的数据库使用模式都与他人不同,每个用户行为中存在的唯一性可以通过适当的事务属性来表示,这有助于构建其行为配置文件[7],并识别攻击者执行的任何恶意尝试. 通常,当入侵者试图通过提交各种非法事务查询来破坏数据库时,通过日志挖掘和犯罪学程序研究的内部入侵检测与防护系统,用于显示和区分普通行为和入侵者的客户端配置文件,从而达到侵入活动的识别[8]. 随着数据价值的增加,数据库系统遭受攻击从未停止,因此数据库入侵检测系统(database intrusion detection system, DIDS)方面的研究不断深入[9].

Rao 等[10]提出一种基于角色访问控制的数据库恶意行为检测方法,设计了基于加权角色的数据依赖性规则挖掘算法,从数据库日志中挖掘出基于加权角色的数据依赖规则,违反数据依赖规则的事务被检测为恶意事务. Elaziz 等[11]提出了增强顺序数据挖掘数据库入侵检测模型,所提出的算法对用户正常历史数据进行挖掘,并对产生的规则进行归并更新,通过训练学习生成异常检测模型,并利用此模型实现基于数据挖掘的异常检测. Wang 等[12]提出了一种基于粗糙概念的多层数据库入侵检测模型,提取计算机数据库的入侵特征,建立粒子群鉴别树进行节点分层处理. 通过不同层次数据库入侵检测的概率操作,实现了多层次、分布式、大型差异数据库的入侵检测. Yi[13]提出了一种利用数据挖掘技术的数据库入侵检测系统,根据相关系统数据提取特定行为特征和规则,利用误用检测和异常检测方法实现入侵检测. 现有数据库入侵

① 收稿日期:2018-05-16
基金项目:中央国家机关支持项目(2011B026).
作者简介:曹德胜(1971-),男,硕士,副教授,主要从事软件工程及数据库研究.

检测系统在保证数据库不受入侵的同时，误报率也会上升.

针对这个问题，本文提出了一种新的数据库入侵检测系统，该系统创新性地将密度聚类技术的点排序识别聚类结构(Ordering Points to Identify Clustering Structure，OPTICS)引入到数据库入侵检测系统，在数据训练阶段，使用OPTICS从用户历史数据库中提取用于构建正常用户配置文件的事务特征. 然而，数据库用户工作职能的转移可能会导致数据库活动出现偏差，这些数据库活动显示为异常值，但不一定是恶意的. 因此，本文系统进一步单独使用多个监督分类器来加强聚类模块的初步结果，学习组件的结合最大限度地减少了数据库所有者因入侵而遭受的损失. 在本文工作中，已经应用了5种不同的监督算法，说明本文系统的可用性和普适性.

1 OPTICS聚类算法理论

OPTICS是一种基于密度的聚类技术，用于发现不同密集区域的聚类，是具有噪声的基于密度聚类方法(Density-Based Spatial Clustering of Applications with Noise，DBSCAN)的扩展，OPTICS的基本思想是：对于簇C_i中的每个对象k，其ε邻域($N_\varepsilon(k)$)中至少存在P个点，其中ε表示半径，P表示创建群集所需的数据点数量. 此外，OPTICS计算数据集中每个数据点的核心距离(dis_c)和可达性距离(dis_r).

可以将对象k的核心距离$dis_c(k)$定义为实例k与其邻域$N_\varepsilon(k)$中对象之间的最小距离，表示为

$$dis_c(k)=\begin{cases}\text{未定义}, & |N_\varepsilon(k)|<P\\ P\_dis(k), & |N_\varepsilon(k)|\geqslant P\end{cases} \tag{1}$$

k到另一个核心对象q对应的可达性距离$dis_r(k)$被定义为使得k从q直接密度可达的最小距离，如果在$N_\varepsilon(k)$中找到至少P个数的实例，则数据点k可以被称为核心点.

$$dis_r(k)=\begin{cases}\text{未定义}, & |N_\varepsilon(k)|<P\\ \max(dis_c(q),\ dis(q,\ k)), & |N_\varepsilon(k)|\geqslant P\end{cases} \tag{2}$$

从式(2)可以推导出点k的本地可达性距离$dis_{lr}(k)$，其可以被描述为与k的P最近邻居的平均可达性距离的倒数.

$$dis_{lr}(k)=\frac{1}{\sum\limits_{o\in N_P(k)} dis_{lr}(k,\ o)/|N_P(k)|} \tag{3}$$

其中，o是k的邻居，$N_P(k)$表示P邻域，偏离集群的点可以看作是异常值. 为了确定对象k是否是离群值，针对每个对象计算局部离群因子(Local Outlier Factor，$\mathrm{LOP}_p(k)$)，其被定义为P最近邻居和k的dis_{lr}的比率平均值.

$$dis_{lr}(k)=\frac{\sum\limits_{o\in N_P(k)}\dfrac{dis_{lr}(o)}{dis_{lr}(k)}}{|N_P(k)|} \tag{4}$$

据观察，位于集群内实例的LOF值接近1. OPTICS算法能够在变化密集的地区识别出有意义的群集，群集方法将类似的数据库访问特征分组到群集中. 本文将OPTICS引入到数据库入侵检测系统中，在数据训练阶段使用OPTICS生成用户配置文件，并根据局部离群因子LOF的值对用户行为进行判断.

2 本文入侵检测系统

本文提出的基于密度聚类数据库入侵检测系统，最初对用户的原始事务数据执行数据预处理，其中所有属性值都被映射为数字. 接下来是应用归一化过程，其中所有属性值都在[0，1]范围内转换，归一化过程很有必要，因为数值高的数据项可能会影响异常计算的结果. 然后，基于密度的聚类技术，即OPTICS被应用于预处理数据集，基于其数据库访问模式中的相似性来构建配置文件簇. 位于集群中的一个事务被标记为真的，而不属于任何群集的事务被传递到不同的监督分类器-朴素贝叶斯(Naive Bayes，NB)、决策树(Decision Tree，DT)、规则归纳(Rule Induction，RI)、K-近邻(k-Nearest Neighbor，k-NN)和径向基函数网络(Radial Basis Function Network，RBFN)进行进一步的研究.

这些分类器通过组合当前事物的信息以及数据库用户的过去行为来检测恶意事物. 在图1中，本文提

出的入侵检测系统事物流包括 2 个阶段：训练阶段和入侵检测阶段.

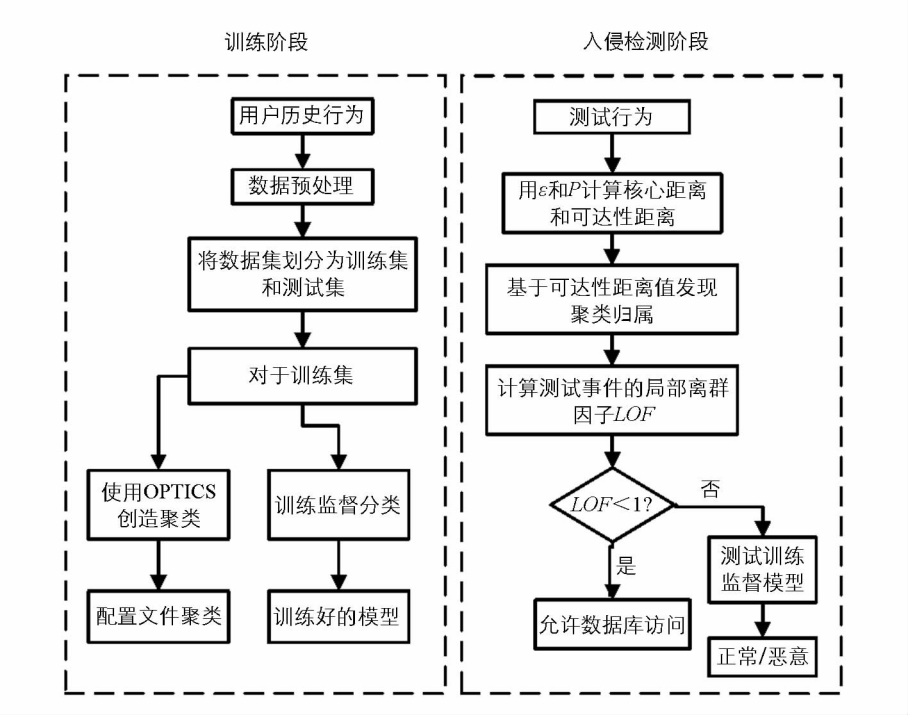

图 1　本文数据库入侵检测系统

2.1　训练阶段

训练阶段讨论与数据库用户原始数据预处理有关的程序，以及用户行为概况构建和 5 种不同监督分类器的训练.

训练事物可以由＜user_id，querytype，tablelist，attlist，timeslot，timegap，loc＞这个属性来表示，其中每个属性的解释见表 1.

表 1　训练事物属性

user_id	用于识别每个用户的唯一标识号
querytype	指定事物请求中涉及的查询类型
tablelist	为数据库中的每个表分配一个唯一的 ID
attlist	表示在特定事务中访问的属性 ID 列表
timeslot	表示在一天中执行事物的时间段，一天 24 h 被划分为每 30 min 的 48 个隙
timegap	描述同一用户在几分钟内连续两次访问数据库之间的时间差
loc	提供了有关事物执行地点的信息，本文已将其映射为数字，例如 1 代表办公室，2 代表家庭.

假设，一个包含两个查询 q_1 和 q_2 的事务已经由具有 user_id ＝ 10 的用户提交到执行某些任务的数据库中，事务中的查询为 q_1：从表 T_1 中选择 x,y，其中 $z=1$；q_2：从表 T_2 中删除，其中 $w=1$.

其中，表 T_1 由属性 x,y 和 z 组成，而表 T_2 由 w 作为其特征，对于 q_1，其访问的属性列表是〈z，x，y〉，而对于 q_2，其访问的属性列表是〈w〉. 在事务中，querytype 为〈SELECT，DELETE〉，tablelist 为〈T_1，T_2〉，attlist 为〈z，x，y，w〉. 将分类值映射到整数后，假设 querytype〈SELECT，DELETE〉＝〈1，4〉，attlist〈z，x，y，w〉＝〈40，23，12，6〉和表格〈T_1，T_2〉＝〈3，6〉. 假设用户在上午 6 点至下午 6 点 30 分(timeslot ＝ 37)之间执行了事物(loc＝1)，并且距离用户最后一次事物的 timegap 为 21 min. 因此，用户 10 的配置文件可以描述为〈10，{1，4}，{3，6}，{40，23，12，6}，37，1，21〉.

为了构建用户配置文件，首先执行事务属性的数据预处理，然后将数据集划分为训练集和测试集. 数据预处理结束后，通过在训练集上应用 OPTICS 算法来构建用户配置文件，其需要两个参数 ε 和 P 作为输入. 根据 OPTICS 算法的式(1) 和式(2) 来计算每个数据点的核心距离(dis_c) 和可达性距离(dis_r) 值，OPTICS 算法以升序方式产生数据点的 dis_c 和 dis_r 值. 通过累积具有相邻 dis_r 值的数据点来形成聚类.

此外，监督分类器的训练是通过将历史训练事物作为输入来完成的，这些分类器分别从数据集中生成各自的学习模型，然后将这些训练好的模型用于异常事务处理以作最终决策.

2.2 入侵检测过程

无论用户何时向数据库提交测试事务，聚类模块使用式(2)计算来自配置文件聚类的 dis_r 值，以确定由最低 dis_r 值决定的聚类归属. 另外，通过使用式(4)计算其局部异常因子(LOF)来确定交易的异常程度. 如果 $LOF<1$，那么事物被标记为真实，并且允许数据库访问. 相反，对于 $LOF\geqslant 1$ 的事务被发现偏离其正常状态，因此经过一系列训练有素的监督分类器进一步处理，以确认数据库访问模式中的偏差. 这些经过训练的分类器模型用于预测不一致事务的最终结果.

3 实验结果与分析

本文使用 Panigrahi 等[14]提出的数据生成程序来生成模拟事物及其标签，以表示真实客户及入侵者的行为. 该模拟器通过使用马尔可夫调制泊松过程(Markov Modulated Poisson Process，MMPP)构建. MMPP 的状态定义和控制来自真实用户和入侵者事务请求的到达率. 此外，3 种高斯分布函数已被用于生成不同的事务属性，以模拟不同类别的真实用户和入侵者. 事务生成在表属性(attlist，tablelist)及事务属性(querytype，timeslot，timegap，loc)的粒度级别上进行控制.

本文采用准确度(Accuracy，Acc)、真正类率(true positive rate，TPR)和负正类率(false positive rate，FPR)来对系统的性能进行验证. Acc 表示正确分类事物的百分比，TPR 表示所识别出的正实例占所有正实例的比例，而 FPR 表示错认为正类的负实例占所有负实例的比例.

训练集和测试集的比例为 7∶3，本文提出的基于密度聚类数据库入侵检测系统的测试由聚类模块开始，该模块以参数 ε 和 P 两个参数作为输入. 为了获得最佳的参数值选择，本文进行了大量实验，得到数据性能随着构建集群数据点数量 P 的增大而变好，而随着 ε 增加而变差. 本文实验中选择了 P 取值为 10，50 和 100，ε 取值为 0 和 0.9，得到不同 ε 和 P 组合的分类性能(表 2).

表 2 不同 ε 和 P 组合的算法分类性能 %

参数组合	性能指标		
	Acc	TPR	FPR
$P=10$，$\varepsilon=0$	56.34	58.79	34.73
$P=10$，$\varepsilon=0.9$	56.24	58.69	35.09
$P=50$，$\varepsilon=0$	58.37	60.12	34.94
$P=50$，$\varepsilon=0.9$	58.27	60.02	35.53
$P=100$，$\varepsilon=0$	62.58	64.3	37.11
$P=100$，$\varepsilon=0.9$	62.48	64.2	38.46

从表 2 中可以看出，当 $P=10$ 时，$\varepsilon=0$ 条件下比 $\varepsilon=0.9$ 条件下性能更佳，此时 Acc 和 TPR 的值都高于 $\varepsilon=0.9$ 产生的性能，而 $FPR=34.73\%$，低于 $\varepsilon=0.9$ 产生的 FPR. 并且随着 P 值越来越大，Acc，TPR 和 FPR 的值都随着升高.

图 2－图 4 给出了本文所提系统使用每个单独的监督分类器(NB，DT，RI，k-NN 和 RBFN)进行入侵检测的性能.

从图 2 中可以看出，在 Acc 性能方面，k-NN 分类器表现最好，其次是 RBFN 分类器、RI 分类器和 NB 分类器，DT 分类器性能最差，只有 90.1%.

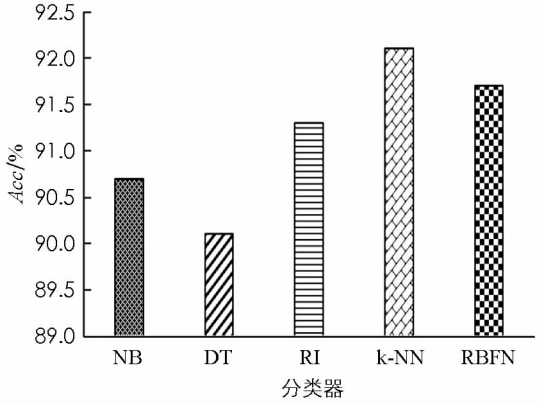

图 2 不同分类器在本文系统中的 Acc 性能

从图 3 中可以看出，在 TPR 性能方面，k-NN 分类器表现最好，其次是 RBFN 分类器、RI 分类器和

DT 分类器，NB 分类器性能最差，只有 90.25%.

从图 4 中可以看出，在 *FPR* 性能方面，NB 分类器表现最好，其次是 RI 分类器、k-NN 分类器和 DT 分类器，RBFN 分类器性能最差，*FPR* 高达 6.87%.

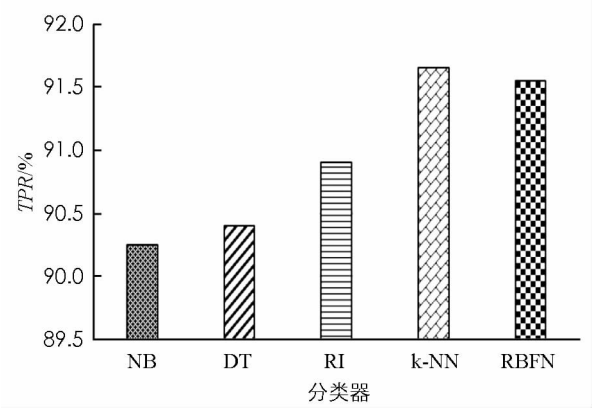

图 3　不同分类器在本文系统中 *TPR* 性能

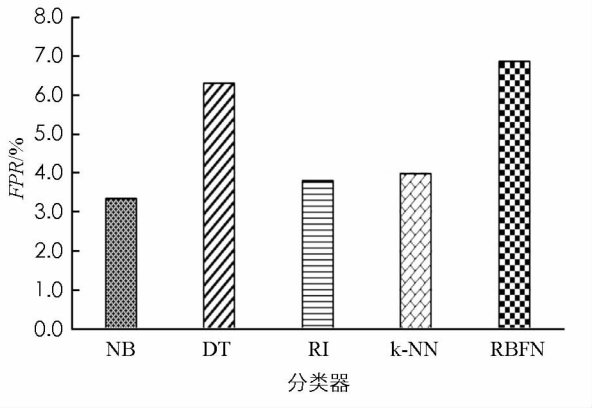

图 4　不同分类器在本文系统中 *FPR* 性能

由图 2－图 4 所示的结果可以得到，k-NN 分类器在 $Acc=92.05\%$ 和 $TPR=91.65\%$ 情况下优于其他分类器，而 NB 分类器在 $FPR=3.35\%$ 情况下具有最低值.

为了验证本文系统性能的优越性，将本文系统与现有系统进行比较，现有系统为 2017 年文献[10]中访问控制启用的数据库恶意检测系统和 2016 年文献[13]中关联规则数据挖掘的数据库入侵检测系统. 分别从 *Acc*，*TPR* 和 *FPR* 3 个方面进行对比，结果见表 3. 通过图 2－图 4 中性能比较，本文系统选择 kNN 作为监督分类器进行入侵检测.

表 3　不同系统的性能比较　　%

方法	*Acc*	*TPR*	*FPR*
文献[10]	91.2	90.1	4.76
文献[13]	90.5	89.2	6.33
本文	92.1	91.7	3.95

从表 3 中数据可以看出，本文数据库入侵检测系统在 *Acc*，*TPR* 和 *FPR* 性能方面都优于其他两种现有系统，说明本文方法的有效性.

4　结　　论

本文提出了一种新的数据库入侵检测系统，该系统引入基于密度的聚类 OPTICS 算法构建数据库用户配置文件，入侵检测方法包括两个阶段：训练和入侵检测. 在训练阶段，对输入数据集的特征进行预处理，并将 OPTICS 聚类建立行为配置文件特征以及监督分类器的训练. 在入侵检测阶段，每个传入事务都由集群模块处理，用于过滤合法模式，并将不一致和错误的事务传递给每个单独受过训练的监督模型以进行最终决策. 使用随机模型进行大规模实验来验证本文系统的有效性，使用的分类器有 NB，DT，RI，k-NN 和 RBFN. 结果表明，本文系统能够使用不同机器学习技术进行入侵检测. 另外，通过与现有数据库入侵检测系统对比，本文系统的性能优于其他系统，说明本文系统的可行性和有效性.

参考文献：

[1] 李　洋，吕家恪. 基于 Hadoop 与 Storm 的日志实时处理系统研究 [J]. 西南师范大学学报(自然科学版)，2017，42(4)：119-126.

[2] 曾　强，缪　力，秦　拯. 面向大数据处理的 Hadoop 与 MongoDB 整合技术研究 [J]. 计算机应用与软件，2016，33(2)：21-24，37.

[3] ASHFAQ R A R，WANG X Z，HUANG J Z，et al. Fuzziness Based Semi-Supervised Learning Approach for Intrusion Detection System [J]. Information Sciences，2017，378：484-497.

[4] 张礼哲，顾兆军，何　波，等. 多源攻击模式图入侵检测方法 [J]. 计算机工程与设计，2016，37(11)：2909-2916.

[5] 陈 虹，万广雪，肖振久. 基于优化数据处理的深度信念网络模型的入侵检测方法[J]. 计算机应用，2017，37(6)：1636-1643，1656.

[6] LAI S F，SU H K，HSIAO W H，et al. Design and Implementation of Cloud Security Defense System with Software Defined Networking Technologies[C]//2016 International Conference on Information and Communication Technology Convergence (ICTC). Jeju：IEEE，2016.

[7] DAWLE Y，NAIK M，VANDE S，et al. Database Security Using Intrusion Detection System[J]. Database，2017，2(3)：1-6.

[8] SURYAWANSHI S S，MULANI T，ZANJURNE S，et al. Database Intrusion Detection and Protection System Using Log Mining and Forensic Analysis[J]. Int J Comput Sci Inf Technol，2015，6：5059-5061.

[9] BUCZAK A L，GUVEN E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection[J]. IEEE Communications Surveys & Tutorials，2016，18(2)：1153-1176.

[10] RAO U P，SINGH N K. Weighted Role Based Data Dependency Approach for Intrusion Detection in Database[J]. IJ Network Security，2017，19(3)：358-370.

[11] ELAZIZ P E A，MOHAMED H K. Database Intrusion Detection Using Sequential Data Mining Approaches[C]//2014 9th International Conference on Computer Engineering & Systems (ICCES). Cairo：IEEE，2014.

[12] WANG N，LI Y，YUAN L M. Simulation on Optimized Intrusion Detection of Multi-Layer，Distributed and Large Differences Database[J]. Applied Mechanics and Materials，2014，556-562：2886-2889.

[13] YI M. On the Research of Force into Computer Database Intrusion Detection Technology[J]. R Risti Iberian Journal on Information Systems & Technologies，2016，18：80-89.

[14] PANIGRAHI S，SURAL S，MAJUMDAR A K. Two-Stage Database Intrusion Detection by Combining Multiple Evidence and Belief Update[J]. Information Systems Frontiers，2013，15(1)：35-53.

On Database Intrusion Detection System Based on Density Clustering

CAO De-sheng

School of Computer Science，North China Institute of Science and Technology，Beijing 065201，China

Abstract：Aiming at the problem of high false positive rate of existing database intrusion detection systems，a database intrusion detection system based on density clustering was proposed in this paper. The intrusion detection system is divided into two parts. ①Data training stage：in this stage，data preprocessing of transaction attributes is executed，and then the data set is divided into training set and testing set. And ordering of points to identify clustering structure (OPTICS) is used to construct the user's normal configuration file；②Intrusion detection stage：each incoming behavior has two states，located within or outside the cluster，and the degree of abnormality of the transaction is determined by its local outlier factor (LOF) value. For LOF<1 behavior allows access to the database，for other behaviors，through the use of different supervised machine learning technology to further verify that the normal/abnormal value，to achieve intrusion detection. The experimental results show that compared with other existing database intrusion detection systems，the performance of this system is better than the other two systems.

Key words：intrusion detection；density clustering；ordering points to identify clustering structure；local outlier factor；supervised learning

责任编辑 夏 娟