

一种处理混合型属性的无监督异常入侵检测方法

郑苗苗, 吉根林

(南京师范大学 数学与计算机科学学院, 江苏 南京 210097)

[摘要] 针对目前入侵检测技术训练时处理类别型数据能力欠缺、误报率高的问题, 提出一种处理混合型属性的无监督异常入侵检测方法, 定义了类别型属性各取值之间的差异度, 使得在对训练集进行无监督学习、生成检测模型过程中, 能够同时有效地处理数值型属性和类别型属性. 理论分析表明所定义的类别型属性值差异度既保留了类别型属性各取值之间的本质特征, 同时也没有改变数据集的原始维数. 实验中采用了网络入侵检测数据集 KDD-CUP-99 来训练模型. 实验结果表明, 采用的混合型属性处理方法进行聚类所建立的入侵检测模型, 与现有方法相比, 检测率高.

[关键词] 入侵检测, 聚类, 混合型属性

[中图分类号] TP311 [文献标识码] A [文章编号] 1672-1292(2008)02-0068-06

An Unsupervised Anomaly Intrusion Detection for the Mixed Attributes

Zheng Miaomiao, Ji Genlin

(School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, China)

Abstract The current intrusion detection techniques can not analyze the attributes composed by categorical and suffer higher false detection rate. In this paper, an effective anomaly detection algorithm based on clustering is proposed to deal with mixed attributes. This algorithm, which gets cluster models by using the clustering algorithm on unlabeled training data, defines the distance between each pair of values in one categorical attribute, can deal with both the numerical and categorical attribute efficiently. Theoretical analysis shows that it holds not only the essence between different values in one categorical attribute, but also the original dimensions of the dataset. At last, experiments on the KDD-CUP-99 data records of network connections show that our method can detect intrusions more efficiently while maintaining a low false positive rate.

Key words intrusion detection, clustering, mixed attributes

按照入侵模型的不同, 入侵检测技术可分为两类^[1]: 误用检测 (Misuse Detection) 和异常检测 (Anomaly Detection). 误用检测能够在含有不同入侵类型的训练集上获取相应的入侵模型, 但不能检测到训练集上不具有的入侵类型. 异常检测则根据训练集的不同分为监督异常检测^[2]和无监督异常检测^[3]. 监督异常检测通过训练有类别标签的数据集来建立检测模型, 能够检测已知攻击. 然而在网络环境中, 要获得足够的标记数据作为训练集很困难. 无监督异常检测方法解决了上述问题, 可以从无类别标签的训练集中训练出未知的入侵, 从而能够识别已知攻击和新型攻击. 无监督异常检测通常采用聚类分析的方法. 传统聚类方法, 如 K-Means, DBSCAN 等, 所处理的数据对象往往仅限于数值类型, 而实际应用中尤其在入侵检测领域, 待处理的数据往往是混合型, 同时包含数值型属性和类别型属性. 通常的解决方法是仅处理数值型属性或者直接转换类别型属性为某一固定数值, 这种方法简单易行, 但对聚类结果的准确性和可靠性有很大影响.

针对目前入侵检测技术训练时处理类别型数据能力欠缺、误报率高的问题, 本文提出了一种处理混合型属性的无监督异常入侵检测方法, 定义了类别型属性各取值之间的差异度, 使得在对训练集进行无监督学习、生成检测模型过程中, 能够同时有效地处理数值型属性和类别型属性. 理论分析表明, 本文定义的类

收稿日期: 2007-09-25

基金项目: 江苏省自然科学基金 (BK2005135) 资助项目.

通讯联系人: 吉根林, 教授, 博士生导师, 研究方向: 数据库与数据挖掘技术、机器学习、XML 技术、入侵检测等. E-mail: glj@njnu.edu.cn

别型属性值差异度既保留了类别型属性各取值之间的本质特征,同时也没有改变数据集的原始维数.实验中采用网络入侵检测数据集 KDD-CUP-99^[4]来训练模型,实验结果表明,采用本文的混合型属性处理方法进行聚类所建立的入侵检测模型,与现有方法相比,检测率高.

1 混合属性处理方法

定义 1 给定一个 d 维数据集 DB , 划分为 k 个簇 w_1, \dots, w_k , 中心点依次为 c_1, \dots, c_k , 其中 $c_j = \frac{1}{n_i} \sum_{j=1}^d x_{ij}$ 是簇 w_i 中数据对象 x_i 的第 j 维属性取值, n_i 是簇 w_i 中数据点总数.

1.1 相关工作

聚类分析中,属性的选择十分重要,直接影响聚类结果的准确性和可靠性.入侵检测数据集往往是混合型数据集,同时包含数值型和类别型属性,如连接传送的字节数、同一端口的连接数等都属于数值型属性,而连接使用的协议类型等则属于类别型属性.在标准化时这两类特征值采用了不同的处理方法.

1.1.1 数值型属性的标准化处理

对于数值型属性而言,不同属性往往具有不同的单位和量纲,其数值的差异可能很大,如果对原始数据不进行预处理的话就有可能产生大数吃小数的问题.如:记录连接时间长度的 duration 属性通常在区间 $[0, +\infty)$ 上取值;记录对于同一服务的连接所占百分比的属性 dst host same srv. rate 通常在区间 $[0, 1]$ 上取值;记录对于同一端口的连接所占百分比的属性 dst host same port rate 通常也在区间 $[0, 1]$ 上取值.显然,属性 duration 掩盖了其余特征,因此要采用标准化和正规化的方法处理数值型属性.通常有以下几种方法^[5]:

总和标准化: $x_{ij} = x_{ij} / \sum_{i=1}^{n_i} x_{ij}$ 通过这种方法得到的新数据 x_{ij} 满足 $\sum_{i=1}^{n_i} x_{ij} = 1$

标准差的标准化: $x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_i}$. 其中,第 j 个特征的均值 $\bar{x}_j = \frac{1}{n_i} \sum_{i=1}^{n_i} x_{ij}$, 标准差 $s_i = \sqrt{\frac{1}{n_i} \sum_{i=1}^{n_i} (x_{ij} - \bar{x}_j)^2}$.

通过这种方法得到的新数据 x_{ij} 的均值 $\bar{x}_j = \frac{1}{n_i} \sum_{i=1}^{n_i} x_{ij} = 0$ 标准差 $s_i = \sqrt{\frac{1}{n_i} \sum_{i=1}^{n_i} (x_{ij} - \bar{x}_j)^2} = 1$

极大值标准化: $x_{ij} = \frac{x_{ij}}{\max\{x_{ij}\}}$. 通过这种方法得到的新数据 $x_{ij} \in [0, 1]$.

极差的标准化: $x_{ij} = \frac{x_{ij} - \min\{x_{ij}\}}{\max\{x_{ij}\} - \min\{x_{ij}\}}$. 通过这种方法得到的新数据 $x_{ij} \in [0, 1]$, 各属性的基点和变化范围都相同.

1.1.2 类别型属性的转化处理

对于类别型属性而言,通常采用直接数值映射方法,也即转换类别型数据为特定数值,如 protocol type 属性有 tcp, udp, icmp 等多种可能的取值,直接数值映射方法将 protocol type 的取值直接映射为一个自然数的集合, tcp 取值为 1, udp 取值为 2, icmp 取值为 3, 依次类推.类别型属性与数值型属性的不同之处在于:数值型属性取值之间有大小关系,而类别型属性各种取值之间只有相同或者不同的关系而没有大小区别,这种方法错误地认为该类型特征之间存在大小关系,对聚类结果的准确性和可靠性有很大影响.文[2]采用了编码映射的方法来处理类别型属性,对于有 s 种不同取值的类别特征,用 s 比特对其进行编码,当且仅当特征取值为第 s_i 种时,其码字中的第 s_i 比特为 1,其余比特为 0.假设属性 protocol type 在数据集中只有 tcp, udp, icmp 3 种取值,文[2]首先将这 3 种取值编码为 001, 010, 100.接下来将 protocol type 属性分割为 3 个属性 protocol type1, protocol type2, protocol type3.当某一记录在该属性上取值为 tcp 时,令 protocol type1=0, protocol type2=0, protocol type3=1;当取值为 udp 时,令 protocol type1=0, protocol type2=1, protocol type3=0.依此类推,将类别型的 protocol type 属性变量转换为多维连续型属性变量.编码映射实际上是将具有多种取值的类别型属性转换为多个具有布尔取值的新属性.该处理方法保留了类别型属性不同取值之间的本质特征,可以确保每条记录之间的同一类别型属性之间距离相等,不会对算法造成误导.但对于取值可能很多的类别型属性来说,完成编码所需要的比特位数会很长,如属性 service

在数据集中可能的取值达 70 种, 对此类属性做编码映射会大幅度增加维数, 计算代价大, 聚类速度慢.

1 2 混合属性处理方法

在不影响聚类速度的情况下, 为了使聚类算法尽可能得到准确可靠的结果, 本文给出定义 2 来计算类别型属性之间的距离. 这种处理方法既保留了类别型属性不同取值之间的本质特征, 同时也不会增加数据集的原始维数.

定义 2 对于类别型属性 p , 设 P 是其在数据集中所有可能取值的集合, $|P|$ 是其所有可能取值的个数, $Num^p_p = \{Num^p_{p_1}, Num^p_{p_2}, \dots, Num^p_{p_{|P|}}\}$ 是所有取值在数据集中重复出现次数的集合, 其中, $Num^p_{p_1} + Num^p_{p_2} + \dots + Num^p_{p_{|P|}} = |DB|$. 如果数据集中任意两条记录 x_i, x_j 在属性 p 上取值分别为 p_m 和 p_n , $1 \leq m, n \leq |P|$ 且 $m \neq n$, 则这两条记录在该属性上的差异 $d_p(x_i, x_j) = |p_m - p_n| = \frac{1}{|P|} + \frac{|Num^p_{p_m} - Num^p_{p_n}|}{Num^p_{p_m} + Num^p_{p_n}}$, 否则 $d_p(x_i, x_j) = 0$.

由定义 2 可看出, $\frac{1}{|P|}$ 是类别型属性 p 所有可能取值之间的共同差异度; $\frac{Num^p_{p_m}}{|DB|}$ 反映了属性 p 取值为 p_m 时的权重, 该取值在数据集中出现的次数越多, 权重则越大; $\frac{|Num^p_{p_m} - Num^p_{p_n}|}{\frac{Num^p_{p_m}}{|DB|} + \frac{Num^p_{p_n}}{|DB|}} = \frac{|Num^p_{p_m} - Num^p_{p_n}|}{Num^p_{p_m} + Num^p_{p_n}}$

[Q 1) 则反映了属性 p 不同取值之间的个别差异度. 表 1 反映了各变量之间的变化关系.

表 1 各变量之间变化关系
Table 1 The relations between each variables

取值在数据集中出现的频度 / 权重		个别差异度	
p_m	p_n	(p_m, p_n)	$d_p(x_i, x_j)$
高 / 大	低 / 小	大	大, $d_p(x_i, x_j) =$ 共同差异度 + 个别差异度 (p_m, p_n)
低 / 小	高 / 大	大	大, $d_p(x_i, x_j) =$ 共同差异度 + 个别差异度 (p_m, p_n)
高 / 大	高 / 大	小	小, $d_p(x_i, x_j) =$ 共同差异度
低 / 小	低 / 小	小	小, $d_p(x_i, x_j) =$ 共同差异度

分析表 1 可以看出, 为了计算数据集中任意两条记录在类别型属性 p 上的差异, 本文的方法不是强制转换类别型属性值为指定数值, 而是利用了各取值之间的内在联系, 并没有改变类别型属性不同取值之间的本质特征; 同时, 由于只是通过数学计算来得到最终的差异距离, 因此也没有增加数据集的原始维数, 不会影响聚类分析的速度.

定义 3 设 d 维数据集 DB 包含 d_N 个数值型属性和 d_C 个类别型属性, $d = d_N + d_C$, 则数据集中任意两条记录 x_i, x_j 的距离为 $d(x_i, x_j) = \sum_{q=1}^{d_N} d_q(x_i, x_j) + \sum_{p=1}^{d_C} d_p(x_i, x_j)$. 其中, $d_q(x_i, x_j)$ 是这两条记录在数值型属性上的距离, 可以采用任意公制距离计算; $d_p(x_i, x_j)$ 则是这两条记录在类别型属性上的距离, 采用定义 2 的方法计算.

定义 4 对于类别型属性 p , 设 w_i 是其在簇 w_i 中所有可能取值的集合, 则中心点 $c_i = \frac{1}{n_i} \sum_{q=1}^{d_N} x_{iq} + \sum_{p=1}^{d_C} p$, 数据集中任一记录 x_j 与中心点 c_i 之间的距离 $d(x_j, c_i) = \sum_{q=1}^{d_N} d_q(x_j, c_i) + \sum_{p=1}^{d_C} d_p(x_j, c_i)$. 其中, $d_q(x_j, c_i)$ 是该记录与中心在数值型属性上的距离, 可以采用任意公制距离计算; $d_p(x_j, c_i) = \sum_{q=1}^{d_N} \left\{ \begin{matrix} 0 & x_{jp} \\ 1 & x_{jp} \end{matrix} \right\} \left\{ \begin{matrix} w_i \\ p \end{matrix} \right\}$ 是该记录与中心 c_i 在类别型属性上取值间的差异.

2 处理混合型属性的无监督异常检测方法

图 1 是处理混合型属性的无监督异常检测系统的体系结构, 由下列 4 个阶段组成:

- (1) 数据预处理阶段: 对未标签的训练集进行预处理;
- (2) 训练阶段: 用聚类算法 K-M eans或 DBSCAN 从处理好的训练集中得到聚簇模型;
- (3) 检测模型建立阶段: 产生若干检测模型;
- (4) 检测阶段: 使用已标记的聚簇模型对测试集进行识别, 检测是否为入侵.

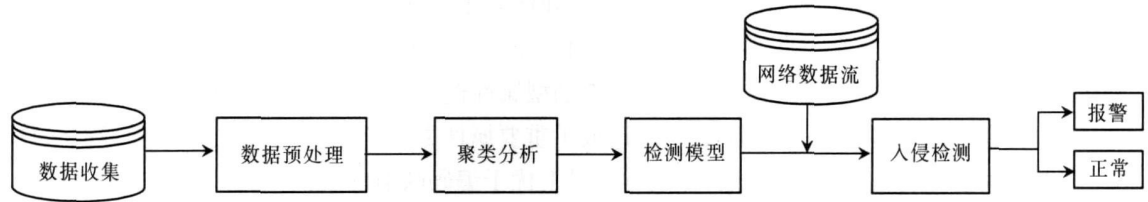


图 1 处理混合型属性的无监督异常检测系统的体系结构
Fig.1 The structure of the unsupervised anomaly intrusion detection system

将聚类算法应用在异常检测中必须基于如下两个前提^[3]: 数据集中绝大部分的数据都是正常数据; 入侵数据在某些属性的取值上偏离正常取值范围. 本文首先对没有标记且仅含少量入侵的训练数据集中的数值属性进行标准化处理, 接下来利用定义 3 定义 4 提出的距离定义对其进行聚类分析 (如 K-M eans 或 DBSCAN). 根据前提一, 利用 DBSCAN 进行聚类分析时, 整个数据集仅被划分为两类, 高密度区域为正常行为, 低密度区域为异常入侵; 而对于 K-M eans 进行聚类分析得到的结果, 本文利用文 [6] 提出的基于簇内样本点数目的类别标记方法标记聚簇: 首先设定一个阈值 N , 当簇内样本点个数占训练集样本总数的比例大于 N 时, 标记该簇正常, 否则为异常. 在实际检测过程中, 对于任意一条新的数据, 测量其与聚类模型中各个聚簇之间的相似度, 找出相似度最大的簇, 查看该簇的标记, 如标记为正常类则认为该数据也为正常, 否则认为其为入侵行为. 实验表明, 利用本文的处理混合型属性的无监督异常检测方法可以有效检测攻击, 做到高检测低误检.

3 实验结果与性能分析

本文通过实验对所提出的处理混合属性的无监督异常入侵检测方法进行性能测试. 实验平台配置为: 100Mb 的局域网; PC 机配置为 Pentium /Intel 2 66G /256MB, W indows XP (Server 版), 80G 硬盘; 开发环境为 Borland JBuilder 2006 Enterprise

为了验证混合属性处理方法的精度, 实验中采用了有类别标签的网络入侵检测数据集 KDD-Cup-99 该数据集来源于 1998DARPA 入侵检测评估程序, 共有 4 900 000 条记录. 每条记录有 41 个特征属性和 1 个真实的分类属性, 其中, 特征属性可细分为 34 个数值型属性、5 个二值属性和 2 个类别型属性; 每条记录都对应着正常模式或某种入侵模式 (DOS R2L, U2R, Probe), 具体分为 1 类正常 22 类攻击. 实验中任意抽取了正常模式和 4 类入侵模式, 构造了 4 个满足两个假设需要的实验数据集, 入侵数据约占数据总数的 1% ~ 1 5%. 实验数据集如表 2 所示.

表 2 数据集
Table 2 Data sets

数据集	记录个数	攻击次数	DOS攻击 /次	R2L 攻击 /次	U2R 攻击 /次	Probe攻击 /次
Data Set 1	10 100	100	25	25	25	25
Data Set 2	10 100	100	50	25	25	0
Data Set 3	10 120	120	30	30	30	30
Data Set 4	10 120	120	30	30	30	30

3 1 模型训练性能

M odha和 Spangler^[7]利用了数据的分类信息来评价聚类结果的好坏, 即当数据有分类信息时, 可认为该分类信息在一定程度上表达了数据的一些内部分布特性. 如果该分类信息没被聚类算法利用, 则可以用

它来评价聚类性能,其度量标准 $Micro\text{-}precision$ 定义为: $Micro\text{-}precision = \frac{1}{n} \sum_{i=1}^k i$ 式中, n 为数据集样本总数; k 为聚类的簇数; i 为聚类的 i 与已知数据集类别对应后,簇 i 中被正确归为相应类别的样本个数. $Micro\text{-}precision$ 的值越大,表示在该数据集上聚类效果越好. 实验中采用该标准来比较算法的精度.

为了验证所提出的混合属性处理方法,本文分别对 34 个数值型属性和所有的 41 个特征属性 (包含 34 个数值型属性、5 个二值属性和 2 个类别型属性)进行了 K-Means 和 DBSCAN 聚类分析,并将聚类结果和真实的分类属性做比较,统计了不同情况下对同一数据集的错分情况,也即 $1 - (Micro\text{-}precision)$ 的值. 其中,处理所有 41 维属性时对类别型属性采用了 3 种处理方法:直接数值映射、编码映射以及本文所提出的基于差异度的计算方法. 实验结果取 10 次实验的平均值. 各算法在处理不同维数时错分率比较如表 3 所示. 分析表 3 可知,只对 34 维属性进行处理和对类别型属性进行直接数值映射的聚类方法错分率高,原因在于未处理所有属性或者处理方式不当,从而造成了重要属性丢失. 本文所提出的采用差异度来计算混合属性的处理方法合理有效,错分率最低,在聚类质量上优于编码映射的类别型属性处理方法. 同时,由于本文的方法保留了数据集的原始属性,聚类速度快于编码映射的方法.

表 3 各聚类算法错分率

Table 3 Comparison of clustering algorithms

各数据 错分率 %	K-Means 算法				DBSCAN 算法			
	34 维	41 维			34 维	41 维		
		数值映射	编码映射	本文方法		数值映射	编码映射	本文方法
Data Set 1	14.75	14.99	8.63	7.24	3.41	7.13	3.327	0.772
Data Set 2	11.09	12.20	6.62	6.11	1.09	5.95	2.23	0.812
Data Set 3	12.16	12.25	7.67	6.56	2.04	6.38	1.94	0.632
Data Set 4	11.78	16.08	7.48	6.70	1.87	7.09	2.61	0.90
All	12.5	13.95	7.64	6.68	2.11	6.67	2.54	0.783

文献 [8] 提出的混合属性距离计算方法 HVDM 和文献 [9] 提出的方法 Minkovdm 是两种常用的混合属性处理方法,效果也较好. 图 2 和图 3 是本文的混合属性处理方法与文献 [8-9] 所提出的方法在不同数据集下结合 K-Means 和 DBSCAN 所得到的效率比较,其结果是取 10 次实验的平均值. 从图中可以看出,本文的方法能达到总体优于文献 [8] 和文献 [9] 的效率,是有效可行的.

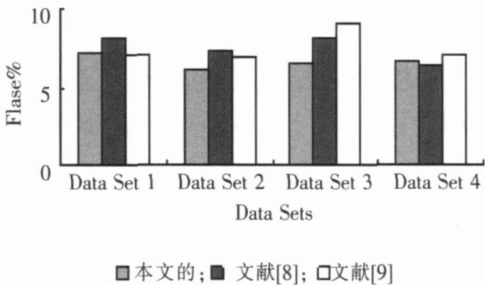


图 2 结合 K-Means 的混合属性处理方法错分比较

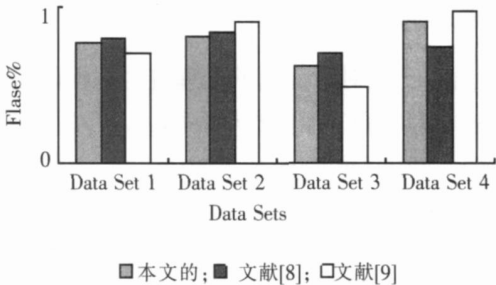


图 3 结合 DBSCAN 的混合属性处理方法错分比较

Fig.2 Comparison of each algorithms based on K-means

Fig.3 Comparison of each algorithms based on DBSCAN

3 2 入侵检测性能

入侵检测系统的性能^[3]主要由检测率 DR 与误检率 FR 两个方面的值来体现. 其中, DR = 检测到的入侵样本数 / 入侵样本总数; FR = 被误报为入侵的正常样本数 / 正常样本总数. 实验中采用表 2 中的数据集两两互为训练集和测试集做交叉验证, 结果如表 4 所示. 从表 4 可见, 在入侵检测中采用处理混合型属性的无监督异常检测方法, 无论四类攻击类型还是攻击类型总集, 所达到的检测率都要比只处理数值型属性所达到的检测率高, 且误检率也相对低. 该方法对 DOS 和 Probe 入侵攻击的检测都取得了较高的检测率, 但对 R2L 入侵的检测效果不理想. 原因在于很多情况下 R2L 攻击采用伪装合法用户身份的方式进行攻击, 其特征与正常数据极为类似, 造成了算法检测的困难. 实际的网络数据中, R2L 攻击所占的比例很小, 该方法的检测结果也会有一定程度的改善.

表 4 入侵检测性能
Table 4 Detection rate and false positive rate

攻击类型	K-Means算法				DBSCAN 算法			
	34 维		41 维		34 维		41 维	
	检测率	误检率	检测率	误检率	检测率	误检率	检测率	误检率
DOS	80		85 18		100		100	
R2L	36 36		45 45		68 18		79	
U2R	63 63	/	70	/	90 9	/	95 5	/
Probe	76 47		80		94 12		94 12	
All	64 31	12 19	70 45	6 4	88 64	1 20	92 5	0 705

4 结论

本文分析了聚类方法在入侵检测中的应用. 针对目前入侵检测技术训练时处理类别型数据能力欠缺、误报率高的问题, 提出了一种处理混合型属性的无监督异常入侵检测方法, 定义了类别型属性各取值之间的差异度, 使得在对训练集进行无监督学习、生成检测模型过程中, 能够同时有效地处理数值型属性和类别型属性. 理论分析表明, 与现有方法相比, 本文定义的类别型属性各取值之间的差异度更为合理, 既保留了类别型属性各取值之间的本质特征, 同时也没有改变数据集的原始维数. 实验结果表明, 通过该方法建立的入侵检测模型能更有效地检测攻击.

[参考文献] (References)

[1] Lee W, Stolfo S J. Data mining framework for building intrusion detection models[C] //Proceedings of the 1999 IEEE Symposium on Security and Privacy. Oakland: IEEE, 1999: 120-132.

[2] Lazarevic A, Ertöz L, Kumar V, et al. A comparative study of anomaly detection schemes in network intrusion detection[C] // Proceedings of the 3rd SIAM International Conference on Data Mining. San Francisco, CA: SIAM, 2003: 1-12.

[3] Portnoy L, Eskin E, Stolfo S J. Intrusion detection with unlabeled data using clustering[C] //Proceedings of the ACM CSS Workshop on Data Mining Applied to Security. Philadelphia, PA: ACM, 2001: 5-8.

[4] The third international knowledge discovery and data mining tools competition dataset KDDCup-99[DB/OL]. [1999-10-28]. <http://kdd-ics.uci.edu/databases/kddcup99/kddcup99.html> [1999].

[5] Jiawei H, Kamber M. Data Mining: Concepts and Techniques[M]. San Francisco: Morgan Kaufmann, 2000: 232-233.

[6] Eskin E, Amodi A, Preneau M, et al. A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data[C] //Proceedings of the Data Mining for Security Applications. Boston: Kluwer Academic Press, 2002: 381-390.

[7] Modha D S, Spangler W S. Feature weighting in k-means clustering[J]. Machine Learning, 2003, 52(3): 217-237.

[8] Wilson D R, Martinez T R. Improved heterogeneous distance functions[J]. Journal of Artificial Intelligence Research, 1997, 6(1): 1-34.

[9] Zhou Z H, Yu Y. Ensembling local learners through multimodal perturbation[J]. IEEE Transactions on Systems, Man and Cybernetics B, 2005, 35(4): 725-735.

[责任编辑: 严海琳]