(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106572108 A
(43)申请公布日 2017.04.19

(21)申请号 201610981248.X

(22)申请日 2016.11.08

(71)申请人 杜少波
地址 455000 河南省安阳市龙安区马投涧镇陈贺驼村岗西中街5号

(72)发明人 杜少波 何文华 杨露 穆肇南 何旭 贾若 李静 袁华 卜艳桃

(74)专利代理机构 贵阳春秋知识产权代理事务所(普通合伙) 52109
代理人 李剑

(51)Int.Cl.
H04L 29/06(2006.01)
G06K 9/62(2006.01)

权利要求书1页 说明书6页

(54)发明名称

一种基于邻域距离的入侵特征选择方法

(57)摘要

本发明公开了一种基于邻域距离的入侵特征选择方法。包括以下步骤:对待聚类数据集采用聚类算法进行聚类,从而获得数据集的类别标签;然后根据类别标签集构成的簇的邻域距离来确定数据集中属性的重要度;最后利用启发搜索来进行特征选择。本发明能够有效地精简入侵数据中的数据特征,提高分类算法的检测效率和检测速度。

CN 106572108 A

1.一种基于邻域距离的入侵特征选择方法，其特征在于：包括以下步骤：对待聚类数据集采用聚类算法进行聚类，从而获得数据集的类别标签；然后根据类别标签集构成的簇的邻域距离来确定数据集中属性的重要度；最后利用启发搜索来进行特征选择。

2.根据权利要求1所述的基于邻域距离的入侵特征选择方法，其特征在于：所述的聚类算法为K-modes聚类算法。

3.根据权利要求1所述的基于邻域距离的入侵特征选择方法，其特征在于：所述属性的重要度的计算公式为：

$$sig(r)=\left|\sum^{k}D_r(X_i,X_j)\right|/C_k^2$$

其中，sig(r)表示属性的重要度，k表示聚类后类别标签集中的分类个数，$D_y(X_i,X_j)$表示X_i和X_j的领域，i＞0，j＞0。

4.根据权利要求1所述的基于邻域距离的入侵特征选择方法，其特征在于：所述的邻域距离的计算公式为：

$$D_p(X,Y)=\frac{\left|(N_p(X)\cup N_p(Y))-(N_p(X)\cap N_p(Y))\right|}{\left|(N_p(X)\cup N_p(Y))\right|}$$

其中，$D_P(X,Y)$表示数据集X和数据集Y属性集$P\subseteq R$上的广义距离，R表示属性的非空有限集合，P为非空有限集合R的子集。

一种基于邻域距离的入侵特征选择方法

技术领域

[0001] 本发明涉及一种入侵特征选择方法，特别是一种基于邻域距离的入侵特征选择方法。

背景技术

[0002] 入侵检测技术是网络安全的一个重要研究方向，它实质上可以归结为安全审计数据的处理。这种处理可以针对网络数据、主机的审记记录或应用程序的日志文件等，目前常用统计技术、分类技术、数据挖掘技术来实现异常行为检测。

[0003] 在入侵检测中，探测器收集到的数据量庞大且提取出来的特征繁多，其中有些特征与检测无关，这些特征一方面降低了分类或聚类的精度，另一方面大大增加了学习及训练的时间和空间复杂度，影响算法运行效率。研究发现，特征选择(Feature Selection,FS)可以在保持原有网络数据信息完整性的基础上去除其中的冗余特征，达到提高系统检测速度的目的。从现有的特征选择算法来看，吴庆涛等提出了一种基于粒子群优化的入侵特征选择算法，该算法通过分析网络入侵数据特征之间的相关性，利用粒子群优化算法在所有特征空间中优化搜索，自主选择有效特征子集，降低数据维度。刘明珍提出了一种二值粒子群优化算法和支持向量机相结合的方法，该算法利用二值粒子群优化算法在特征空间中进行全局搜索，选择最优特征集进行分类。张宗飞提出一种基于量子进化算法的网络入侵检测特征选择算法，该算法将量子进化算法应用于网络入侵检测的特征选择；林冬茂等提出了一种基于无监督免疫优化分层的网络入侵检测算法，该算法将免疫网络与分层聚类方法相结合，达到检测网络入侵的目的。这些特征选择方法都是基于人工智能算法，由于人工智能算法需要一些参数设置，因此参数设置是否合理将直接影响检测算法的性能。

发明内容

[0004] 本发明的目的在于，提供一种基于邻域距离的入侵特征选择方法。它能够有效地精简入侵数据中的数据特征，提高分类算法的检测效率和检测速度。

[0005] 本发明的技术方案：一种基于邻域距离的入侵特征选择方法，包括以下步骤：对待聚类数据集采用聚类算法进行聚类，从而获得数据集的类别标签；然后根据类别标签集构成的簇的邻域距离来确定数据集中属性的重要度；最后利用启发搜索来进行特征选择。

[0006] 前述的基于邻域距离的入侵特征选择方法中，所述的聚类算法为K-modes聚类算法。

[0007] 前述的基于邻域距离的入侵特征选择方法中，所述属性的重要度的计算公式为：

[0008] $$sig(r)=\left|\sum^{k}D_r(X_i,X_j)\right|/C_k^2$$

[0009] 其中，sig(r)表示属性的重要度，k表示聚类后类别标签集中的分类个数，$D_y(X_i,X_j)$表示X_i和X_j的领域，i＞0，j＞0。

[0010] 前述的基于邻域距离的入侵特征选择方法中，所述的邻域距离的计算公式为：

[0011] $$D_p(X,Y)=\frac{\left|(N_p(X)\cup N_p(Y))-(N_p(X)\cap N_p(Y))\right|}{\left|(N_p(X)\cup N_p(Y))\right|}$$

[0012] 其中，$D_P(X,Y)$表示数据集X和数据集Y属性集$P\subseteq R$上的广义距离，R表示属性的非空有限集合，P为非空有限集合R的子集。

[0013] 申请人对本发明进行了大量的研究，如下：

[0014] 实验例1：通过比较具有41个特征的入侵检测模型和经过特征选择后的特征子集的模型在检测率和检测时间上的性能。首先对随机抽样后的数据集应用本文特征选择方法，选择具有最大属性重要度的特征，得到相应的特征子集。然后在训练集上分别对具有41个属性特征和经过特征选择方法选取的特征子集上建立入侵检测模型。

[0015] 实验例2：通过比较本文特征选择方法与遗传算法(Genetic Algorithm,GA)和Relief算法的入侵检测模型在检测率和检测时间上的性能。首先对经过抽样的数据集采用本文特征选择方法，得到相应的特征子集，并应用得到的特征子集建立入侵检测模型，然后用遗传算法和Relief算法对相同训练数据进行特征选取。

[0016] 实验例3：使用SVM作为分类器，通过比较实验例2中的三个特征选择方法获得的特征子集，做为样本分类时所产生的分类错误率。

[0017] 实验结果分析

[0018] 对随机抽样后的训练数据集应用本文特征选择方法，得到如表1所示特征子集。

[0019] 表1 本文特征选择方法得到的特征子集

[0020]

攻击类型	特征子集
Dos	dos,protocol_type,dst_bytes,service,count,dst_host_same_src_rate
U2R	U2R,protocol_type, service, root_shell, dst_host_count,dst_host_srv_rate
R2L	R2L, duration, service, dst_bytes,dst_host_count
PROBE	probe,protocol_type,duration, service,src_bytes,logged_in,hot,count, dst_host_diff_src_rate
NORMAL	protocol_type,service, count,src_bytes, dst_host_count

[0021] 从表1中的数据可以看出经过特征选择后，不同的攻击类型将会得到不同的特征子集。针对每一种攻击类型，分别对未经过特征选择的数据集和应用本文方法选择的特征

子集上建立入侵检测模型，通过比较这两种入侵检测模型在检测率和检测时间方面的性能。结果如表2所示。

[0022] 表2 特征选择前后检测率和检测时间结果对比

攻击类型	分类准确率		检测时间（sec）	
	所有特征	特征子集	所有特征	特征子集
Dos	84.4%	90.6%	1.10	0.33
U2R	86.2%	92.7%	0.92	0.19
R2L	83.7%	92.2%	1.09	0.35
PROBE	88.5%	93.3%	1.15	0.46
NORMAL	87.3%	96.5%	1.03	0.32

[0023]

[0024] 从表2数据中可以看出，本文的特征选择方法在检测准确率和检测时间方面的性能明显优于未经过特征选择的入侵检测模型。在随机抽样后的训练数据集上，分别用本文方法、遗传算法和Relief算法，实验结果比对如表3所示。

[0025] 表3 本文方法、遗传算法和Relief算法结果对比

[0026]

攻击类型	本文方法属性个数	分类准确率			检测时间（sec）		
		本文方法	遗传算法	Relief	本文方法	遗传算法	Relief
Dos	6	96.3%	96.8%	95.4%	0.28	0.33	0.43
U2R	6	94.8%	95.3%	95.3%	0.19	0.25	0.35
R2L	5	97.2%	97.9%	94.7%	0.31	0.39	0.49
PROBE	9	96.5%	97.3%	96.9%	0.36	0.45	0.51

[0027]

NORMAL	5	97.7%	97.6%	95.8%	0.29	0.35	0.46

[0028] 最后，使用SVM作为分类器，将本文方法、遗传算法和Relief算法对经过特征选择后的训练数据集做为分类样本，结果如表4所示。

[0029] 表4 SVM分类器下本文方法、遗传算法和Relief算法结果对比

[0030]

	分类准确率			检测时间（sec）		
数据集	本文方法	遗传算法	Relief	本文方法	遗传算法	Relief
训练子集	96.9%	97.2%	95.4%	0.29	0.38	0.43
测试子集	96.3%	96.7%	96.2%	0.46	0.59	0.55

[0031] 从表3中可以得出，本文特征选择方法在分类准确率方面与遗传算法相比差别不大，而与Relief算法相比有较高的准确率；在检测时间上本文特征选择方法要优于遗传算法和Relief算法。

[0032] 综合以上三个实验可以得出，本文特征选择方法，可以有效的降低高维数据的维度；在经过特征选择的特征子集上建立入侵检测模型在检测率和检测时间上要优于未经过特征选择的入侵检测模型；通过于遗传算法和Relief算法比较，可以看出本文特征选择方法能够在保证较高的检测率的同时，降低特征选择的时间复杂度，能有效地缩短算法的运行时间。

[0033] 与现有技术相比，本发明通过对待聚类数据集采用聚类算法进行聚类，从而获得数据集的类别标签，然后根据类别标签集构成的簇的邻域距离来确定数据集中属性的重要度，能够有效地精简入侵数据中的数据特征，提高分类算法的检测效率和检测速度；针对高维和数据量大的数据集，可以通过该方法得到具有聚类性质的有效特征子集，在得到的特征子集上可以对数据集进行重新聚类分析，有效降低了因高维复杂数据集导致计算耗时高的问题；启发式搜索算法以空集为起点，每次加入使当前特征子集重要度最大的属性，直到特征子集重要度不发生变化时停止搜索，这样可以确保重要的属性优先加入到特征子集中，从而不会丢失重要的特征，这样选出的特征子集作为一个整体能够保持原始数据的分类能力，同时去除了与聚类无关的冗余属性和分类能力差的属性。

具体实施方式

[0034] 实施例。一种基于邻域距离的入侵特征选择方法，包括以下步骤：对待聚类数据集采用聚类算法进行聚类，从而获得数据集的类别标签；然后根据类别标签集构成的簇的邻域距离来确定数据集中属性的重要度；最后利用启发搜索来进行特征选择。所述的聚类算法为K-modes聚类算法。所述属性的重要度的计算公式为：

[0035] $$sig(r)=\left|\sum_{k}^{k}D_r(X_i,X_j)\right|/C_k^2$$

[0036] 其中，sig(r)表示属性的重要度，k表示聚类后类别标签集中的分类个数，$D_y(X_i,X_j)$表示X_i和X_j的领域，i＞0，j＞0；i和j的取值视集合大小而定。

[0037] 所述的邻域距离的计算公式为：

[0038] $$D_p(X,Y)=\frac{\left|(N_p(X)\cup N_p(Y))-(N_p(X)\cap N_p(Y))\right|}{\left|(N_p(X)\cup N_p(Y))\right|}$$

[0039] 其中，$D_P(X,Y)$表示数据集X和数据集Y属性集$P\subseteq R$上的广义距离，数据集之间的

距离越大，则该值就越大；R表示属性的非空有限集合；P为非空有限集合R的子集。

[0040] 邻域的基本概念，定义1假设S＝(U,R)为一个信息系统，其中U是一个对象的非空有限集合，称为论域；R表示属性的非空有限集合；对任意r∈R有r:U→V_r，其中V_r称为属性r的值域；对任意r∈R,x∈U存在f(x,r)∈V_r，其中f(x,r)是一个信息函数，它对论域中对象的每一个属性赋予一个信息值。

[0041] 定义2假设S＝(U,R)为一个信息系统，对于任意的x∈U，x在属性集$P \subseteq R$上的邻域表示为：

[0042] $N_p(x) = \{y \mid f(x,r) = f(y,r), y \in U\}$

[0043] $X \subseteq U$,X在属性集$P \subseteq R$上的邻域表示为：

[0044] $$N_p(X) = (y \mid \forall x \in X, f(x,r) = f(y,r), y \in U)$$

[0045] 定义3给定一个信息系统S＝(U,R)，经聚类后将U划分为k个分类：$C_1, C_2, \cdots, C_k$，$\forall P \subseteq R$，定义类C关于P的邻域表示为：

[0046] $N_p(C) = \{N_p(C_1), N_p(C_2), \cdots, N_p(C_k)\}$

[0047] 特征选择算法

[0048] 定义4给定一个信息系统$S = (U,R), X \subseteq U, Y \subseteq U$，定义X,Y在属性集$P \subseteq R$上的邻域距离为：

[0049] $$D_p(X,Y) = \frac{\left|(N_p(X) \cup N_p(Y)) - (N_p(X) \cap N_p(Y))\right|}{\left|(N_p(X) \cup N_p(Y))\right|}$$

[0050] 其中，D(X,Y)可以看作为数据集X和数据集Y在属性集$P \subseteq R$上的广义距离，数据集之间的距离越大，则D(X,Y)的值也就越大。

[0051] 定义5给定一个信息系统S＝(U,R)，经过聚类后将U划分成为k个分类：$C_1, C_2, \cdots, C_k$，属性r∈R的属性重要度为：

[0052] $$sig(r) = \left(\sum_{\substack{i=1 \\ j>i}}^{k} D_r(X_i, X_j) \right) / C_k^2$$

[0053] 定义4、定义5主要表达不同特征在不同类之间的区分能力。一个具有聚类性质的特征是指该特征可以使类内对象之间的距离尽可能小，而使不同类间的距离尽可能大。也就是说一个具有聚类性质的特征对于不同类之间应该具有较大的区分度。寻找具有聚类性质的特征，可以得到使类内对象之间的距离最小、而不同类之间的距离最大的特征子集。

[0054] 算法描述

[0055] 为了实现对待聚类数据集进行特征子集选择，当数据集具有m个属性时，计算所有的特征子集的分类能力，需要测试2^m-1个特征子集，因此对高维和数据量大的数据集，计算耗时巨大。因此，可以利用启发式搜索算法来解决该问题，该启发式搜索算法以空集为起点，每次加入使当前特征子集重要度最大的属性，直到特征子集重要度不发生变化时停止搜索。这样可以确保重要的属性优先加入到特征子集中，从而不会丢失重要的特征。这样选出的特征子集作为一个整体能够保持原始数据的分类能力，同时去除了与聚类无关的冗余属性和分类能力差的属性。

[0056] 算法步骤如下：

[0057] Step1：在待聚类数据集上随机选取k个初始中心点，并进行K-modes聚类，得到类别标签集C，时间复杂度为：O(knmt)；

[0058] Step2：使用式(5)计算属性的重要度，时间复杂度为$O(kmC_k^2)$；

[0059] Step3：$\varnothing \rightarrow FS$；

[0060] Step4：对任意属性$r_i \in R$-特征子集，计算sig(FS∪{r_i})，时间复杂度为$O(km^2C_k^2)$；

[0061] Step5：选择r_i，满足max(sig(FS∪{r_i}))；

[0062] Step6：如果max(sig(FS∪{r_i}))＞0

[0063] FS∪{r_i}→FS，跳转到Step4；否则，结束输出特征子集，时间复杂度为$O(km^2C_k^2m!)$；

[0064] 通过算法分析可知，该算法总的时间复杂度为：$O(knmt+kmC_k^2+O(km^2C_k^2)+O(km^2C_k^2m!))$。

[0065] 针对高维和数据量大的数据集，可以通过该算法得到具有聚类性质的有效特征子集，在得到的特征子集上可以对数据集进行重新聚类分析，有效降低了因高维复杂数据集导致计算耗时高的问题。