文章编号:1000－2367(2017)05－0101－07 DOI:10.16366/j.cnki.1000－2367.2017.05.017

一种基于可拓距的特征变换方法及其在网络入侵检测中的应用

徐慧,刘翔,方策,宗欣露

(湖北工业大学 计算机学院,武汉 430068)

摘 要:作为识别攻击或异常行为以保护网络安全的重要步骤之一,网络入侵检测常常与数据挖掘或机器学习技术结合应用.如今,随着网络数据的爆炸性增长,传统的入侵检测技术面临着海量数据检测处理的问题,现有入侵检测系统往往难以同时满足实时性和有效性的需求.本文尝试将可拓学中的可拓距概念引入网络入侵检测研究中,提出了一种基于可拓距的特征变换方法,将数据点的原特征映射为簇外中心距和簇内可拓距这两大部分,根据原始数据多维特征生成新的特征,以达到特征降维的目的,旨在同时满足网络入侵检测系统的实时性和有效性的需求.本文使用 KDD CUP 99 作为仿真数据集测试所提出的基于可拓距的方法在网络入侵检测特征变换中的应用效果.实验结果表明,较之传统的 KNN 算法,基于可拓距的方法明显地减少了检测时间,而同时其检测率的下降可以控制在 1%之内,具有较好的时效性优势.

关键词:网络入侵检测;特征变换;可拓学;簇外中心距;簇内可拓距

中图分类号:TP393.08 **文献标志码**:A

随着网络技术不断发展,互联网在军事、医疗、金融等领域的应用越来越广泛,网络安全愈来愈得到重视.为了能及时地发现攻击行为,从而将其所产生的损失降到最低,人们对网络入侵检测系统的实效性提出了更高的要求.传统的检测方法虽然在检测率等方面有着较好的表现,但是需要耗费大量的时间和系统资源.并且,随着技术的发展,许多新的网络攻击方式不断产生,训练集的数量需要不断增加.这将极大地增加运算的时间,难以满足入侵检测系统的实时性要求.

针对上述问题,国内外学者展开了相关研究,提出了一些基于特征变换的降维方法.文献[1]提出了一种基于三角形的混合入侵检测方法 TANN,该方法将一个样本与两个聚类中心进行变换后形成该样本的一个新特征.文献[2]提出了一种基于最邻点与簇中心的特征变换方法 CANN.该方法主要计算 2 个距离.一个是样本点与其最邻近点的距离,另一个是样本点与所有聚类中心的距离.通过计算这 2 个距离之和得到一个新的特征.用这个新的特征取代原数据的原始特征.文献[3]提出了一种基于簇中心距离和的特征提取方法 DSFE.该方法使用多个簇中心的距离和来代替数据样本的原始特征信息.文献[4]提出了一种基于密度、聚类中心以及最邻近点的特征降维方法 DCNN.该方法可以将原数据集中的数据维度降低到二维.

上述适用于入侵检测的特征变换方法都有变换后新数据的特征维数较低、新特征的区分能力较强等特点,但是对于处于同一类的数据的区分并不是很明显.本文通过引入可拓学中的可拓距,提出一种基于可拓距的特征变换(Feature Transformation Based on Extension Distance,FTBED)方法,该方法的基本思路:首先运用聚类算法得到聚类中心;然后根据聚类的结果将数据点的原特征映射为描述数据点簇内和簇外这两大部分特征,既可以做到不同类数据的区分,也能对处于同一类的数据进行区分;最终将 FTBED 算法应用于网络入侵检测的特征变换中.

收稿日期:2017-04-23;**修回日期**:2017-06-19.

基金项目:国家自然科学基金(61602162;61440024;61502155);湖北工业大学博士科研启动基金计划项目(BSQD12029).

作者简介(通信作者):徐慧(1983－),女,湖北武汉人,湖北工业大学副教授,博士,研究方向为网络与服务管理,E-mail:xuhui@mail.hbut.edu.cn.

1 网络入侵检测与可拓学

1.1 k-means 算法

k-means 算法是网络入侵检测领域中运用最广泛的聚类算法. k-means 算法具有实现简单、计算复杂度低的优点,其主要的思想是根据数据点间的相似度将所有的数据点划分到 k 个不同的簇当中. k-means 算法的流程如下所示:

输入 含有 n 个数据点的数据集 D,聚类的个数 k.

步骤 1 从数据集中随机选取 k 个数据点作为初始的聚类中心点.

步骤 2 遍历计算余下所有的数据点到 k 个初始聚类中心点的欧式距离,并将其划分到最近的聚类当中.

步骤 3 分别计算 k 个类中所有的数据点各个维度的算术平均值,得到 k 个新的聚类中心点.

步骤 4 重复步骤 2 和步骤 3,直到满足停止条件为止.

步骤 5 输出聚类结果.

如上述流程所示,k-means 算法步骤 4 采用的停止条件通常是关与类内数据点的平方误差和以及设定的最大迭代次数. 平方误差和是指数据集中的所有数据点与其所在的类的簇中心的平方误差之和

$$H = \sum_{i-1}^{k} \sum_{x \in c_i} (x - x_i)^2, \tag{1}$$

其中,x 是指样本数据点,x_i 是指第 i 个簇的簇心. 通常,当 H 的取值越小时,聚类结果的质量越好.

聚类算法的目的是为了使相似度高的对象分到同一个类中,而差别很大的对象分到不同的类中. 因此,无法做到对同一类中数据点进行区分. 为了描述类内数据点的区分程度,本文引入了可拓学中的可拓距概念.

1.2 可拓距的引入

可拓学是由中国学者 1983 年提出的一门原创性横断学科. 它以形式化的模型,讨论事物的拓展性以及开拓创新的方法,并将其用于解决矛盾问题[5]. 如今,参与研究和应用可拓学的国内外科技工作者也越来越多,可拓学已经广泛应用于多个领域. 文献[6]将可拓学应用于数据挖掘中建立了以可拓集和关联函数为核心的可拓聚类方法. 文献[7]将可拓距应用于起重机产品的配置方法设计中,通过将将相似度标准差与专家经验进行结合进行动态权重分配,使检索特征更好地反映出配置结果. 文献[8]将可拓识别方法与无人车识别障碍物相结合,生成了符合识别要求的策略. 文献[9]将 Ada Boost 算法和可拓学理论相融合,提出了可拓 Ada Boost 算法,该算法在适用问题中使分类结果更准确.

可拓学通过可拓距的概念描述类内事物的区别,在点与区间的距离描述上有很大优势,更符合实际情况. 本文考虑将引入可拓距应用于网络入侵检测的特征变换中. 可拓距对经典数学中点与区间距离这一概念进行了拓展. 在经典数学中,对于存在于区间内的点认为其与区间的距离为 0. 而在可拓学中,可以用可拓距的值来描述不同点在区间内的不同位置[10]. 对于区间 $U = (a,b)$,有实轴上任意一点 $x \in (a,b)$,点 x 与区间 U 的可拓距

$$\rho(x,U) = \left| x - \frac{a+b}{2} \right| - \frac{b-a}{2}. \tag{2}$$

2 基于可拓距的特征变换方法

2.1 核心思想

当点在区间内时,经典数学中认为点与区间的距离都为 0,而在可拓集合中,利用可拓距的概念,就可以根据可拓距的值的不同来描述点在区间内的位置的不同. 可拓距的概念对点与区间的位置关系的描述,使人们从“类内即为同”发展到类内有程度区别的定量描述[10].

考虑到可拓距可以将点与区间的位置关系用定量的形式精确描述,本文将可拓距引入到特征变换研究

中,提出了一种基于可拓距的特征变换方法.该特征变换方法将数据点的原特征映射为 2 大部分:

1)描述数据点与其他簇差异的特征,即簇外中心距;

2)描述数据点与簇内数据点差异的特征,即簇内可拓距.

本文正是通过计算簇内可拓距来描述数据点与簇内数据点差异,进而生成新数据点的第二部分特征.图 1给出该特征变换方法的核心思想.

如图 1 所示,基于可拓距的特征变换方法主要可以分为以下 3 个阶段.

阶段 1　对于任意的一个给定的 m 维数据集 D,应用聚类算法将数据集 D 中的数据点聚类为 k 个簇,提取出簇中心点 $c_1,c_2,c_3,\cdots,c_k$.

阶段 2　根据提取出的簇中心,遍历各数据点,分别计算其簇外中心距 j_w 和簇内可拓距 j_n.

阶段 3　利用簇外中心距 j_w 和簇内可拓距 j_n 将原数据集 D 转换为一个新的 $k+1$ 维数据集 D_1,其中,D_1 中的数据点特征由 k 个簇外中心距离和 1 个簇内可拓距构成.

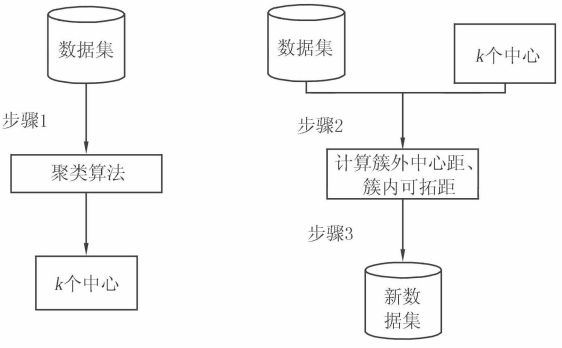

图1　基于可拓距的特征变换方法的核心思想

2.2　算法流程

依据如图 1 所示的方法核心思想,基于可拓距的特征变换方法的算法流程如下所示.

输入　含有 n 个数据点的 m 维数据集 D,指定的聚类数量 k.

步骤 1　从数据集中随机选取初始的 k 个数据点作为聚类中心点.

步骤 2　对数据集 D 中的每个数据点 x_i,计算其到各个聚类中心点的距离,将该数据点归类到距离其最近的一个聚类当中.

步骤 3　对所得到的 k 个聚类,重新计算每个聚类的聚类中心点.

步骤 4　将新得到的聚类中心点和原来的聚类中心点比较,若聚类中心点没有变化,则执行步骤 5;若聚类中心点出现变化,则再次执行步骤 3.

步骤 5　获取 k 个聚类中心 $c_1,c_2,c_3,\cdots,c_k$ 以及 k 个簇 $s_1,s_2,s_3,\cdots,s_k$,分别计算各数据点的簇外中心距离 j_w 和簇内距 j_n.

步骤 6　依次将 k 个簇 $s_1,s_2,s_3,\cdots,s_k$ 中的 m 维数据点转化为 $k+1$ 维的新数据点,得到降维后的数据集 D_1.

如上所述,基于可拓距的特征变换方法将数据点的原特征映射为簇外中心距和簇内可拓距这 2 大部分,下面对上述流程的步骤 5 中这 2 类距离计算展开说明.

2.2.1　簇外中心距计算

新的 $k+1$ 个特征值有 2 部分,一部分是原数据点与 k 个簇中心的距离,另一部分是数据点与自身所在簇中心的簇内可拓距.在获取簇中心之后,便可以计算每个数据点的簇外中心距与簇内可拓距.簇外中心距离 j_w 是根据原 m 维数据点和由聚类算法得到的 k 个簇中心 $c_1,c_2,c_3,\cdots,c_k$,来生成 k 个新的特征向量用以替换原数据点中的特征向量,从而得到降维后的新的特征点.为了计算样本在特征空间中的距离,本文采用欧氏距离函数.对于数据点 $x_i=(x_{i1},x_{i2},\cdots,x_{im})$ 和簇中心 $c_j=(c_{j1},c_{j2},\cdots,c_{jm})$ 的距离

$$d(x_i,c_j)=\sqrt{|x_{i1}-c_{j1}|^2+|x_{i2}-c_{j2}|^2+\cdots+|x_{im}-c_{jm}|^2},\tag{3}$$

其中 x_{im} 代表数据集中第 i 个数据点的第 m 维特征向量的取值,c_{jm} 代表 k 个簇中心中第 j 个簇中心的第 m 维特征向量的值.

如图 2 所示,在第 k 个簇中的数据点 x_i 与 k 个簇的簇心距离分别为 $d(x_i,c_1),d(x_i,c_2),d(x_i,c_3),\cdots,d(x_i,c_k)$,分别记作 $j_{w1},j_{w2},j_{w3},\cdots,j_{wk}$.降维后的特征点 x_i',其前 k 个特征可以记为 $x_i'(j_{w1},j_{w2},\cdots,j_{wk})$.

2.2.2　簇内可拓距的计算

首先要筛选出距离簇心 c 最邻近的点 n_1 以及距离簇心最远的点 n_2.可以通过遍历计算每个簇中其他数

据点与簇中心的距离,选取出与簇中心距离最短的点作为最邻近点 n_1,与簇中心距离最长的点作为最远点 n_2. 如图 3 所示,n_1 和 n_2 即为计算可拓距所选取的参考点,而簇心与这两点的距离 $d(c,n_1)$ 以及 $d(c,n_2)$ 作为参考区间的值,分别记为 a 和 b,进而有可拓距的参考区间为 $X=[a,b]$. 对于簇内任意一点 x_i,其簇内可拓距

$$j_n=\rho(x_i,X)=\rho(d(c,x_i),X)=\left|d(c,x_i)-\frac{a+b}{2}\right|-\frac{b-a}{2}. \tag{4}$$

根据公式(3)和(4)所示的簇外中心距和簇内可拓距的计算方法,可以将原数据点转化为一个新的 $k+1$ 维数据点. 新数据点的特征由以下 2 部分构成:

1) 有 k 个特征,分别为数据点与 k 个簇中心的距离值;

2) 则为簇内可拓距 j_n.

根据上述过程可以将原 m 维的数据点转化为新的 $k+1$ 维数据点,记为 $x_i'(j_{w1},j_{w2},j_{w3},\cdots,j_{wk},j_n)$. 对于原数据点中所带有的类别标签则不做改变,在新的数据点中保留.

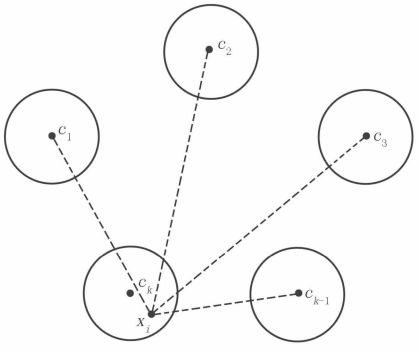

图2 簇外中心距

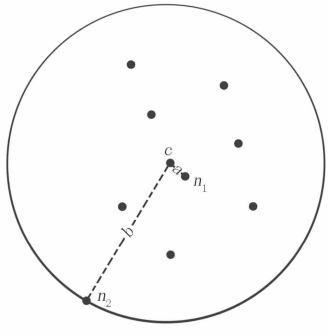

图3 可拓距参考点的选取

3 在网络入侵检测中的应用

3.1 实验流程设计

为了获得簇中心,可以使用一种聚类算法对数据集进行聚类. 本文考虑的是在入侵检测领域当中使用最为广泛的 k-means 算法. 对于 k-means 算法,k 的取值是一个关键问题,而在网络入侵检测领域中,选用某个数据集时往往会得知数据集中样本的类别数. 本文选择网络入侵检测领域中应用最为广泛的 KDD CUP 99 数据集作为实验数据集. 在 KDD CUP 99 数据集中样本类别数为 5. 因此,将 k-means 算法的参数 k 值设置为 5,则根据 k-means 算法的聚类结果,可以得到 5 个簇中心 c_1,c_2,c_3,c_4,c_5.

具体而言,本文选择 KDD CUP 99 数据集的子集作为实验数据,其中包含 494 021 条连接记录,该数据子集中的每一条连接记录都可以归类为 NORMAL,PROBING,DOS,U2R,R2L 这 5 种类别之一. 每一条记录都包含有 41 个特征,其中 38 个为数值型特征,3 个为字符型特征[11-12]. 在开始阶段,对数据集进行预处理,将 3 个字符型特征映射为数值型特征,并对所有的特征做最小-最大规范化处理,

$$V'=\frac{\nu-m_A}{M_A-m_A}(M_A'-m_A')+m_A', \tag{5}$$

其中 m_A' 取值为 0,M_A' 取值为 1.

实验流程为:首先分别按 3 种不同的比例(1.5%、5.5%、10.4%)对 KDD CUP 99 数据集进行抽样,然后使用 k-means 算法对抽样后的数据进行聚类,根据聚类的结果将原始数据集转化为新的实验数据集,接着按比例进行训练集和测试集的区分,其中训练集占 20%,测试集占 80%. 最后,使用 KNN 作为分类算法,得到各类数据的分类正确率.

根据多次试验发现 KNN 算法的 k 值取值为 10 时效果较好,因为过大的 k 会使在对样本较少的类别如 U2R 进行分类时,正确率较低,甚至出现全部错分的情况,而较小的 k 则使算法正确率有所降低.

3.2　实验结果分析

表 1 给出这两种算法的准确率对比.

表 1　KNN 算法与所提出的 FTBED 算法的 3 组准确率对比

数据量/条	准确率/%	
	KNN 算法	FTBED 算法
7615	96.109	95.207
27 241	98.901	98.411
51 440	99.258	98.900

如表 1 所显示的，尽管 KNN 算法的准确率要略高于 FTBED，但两者准确率的差值均在 1%之内. 图 4 为 2 个算法的耗时对比图，而图 5 则为两个算法对应的准确率差值变化图.

一方面，如图 4 所示，FTBED 算法耗时要比 KNN 少. 并且随着样本数量的增加，KNN 算法与 FTBED 算法间的耗时差距越来越大. 另一方面，如图 5 所示，随着数据量的增加，准确率的差值在不断地减小，说明两个算法在准确率方面并没有很大的差别.

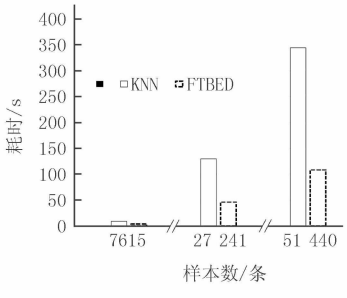

图4　KNN算法与所提出的FTBED算法的耗时对比图

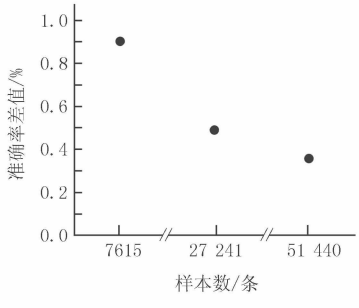

图5　KNN算法与所提出的FTBED算法的耗时对比图

表 2 和表 3 分别表示 KNN 算法与所提出的 FTBED 算法在数据量为 7615 条时对各类样本的准确率.

表 2　数据量为 7615 条时 KNN 算法对各类样本的准确率

实际类别	预测结果/条					准确率/%
	NORMAL	PROBING	DOS	U2R	R2L	
NORMAL	724	17	1	0	30	93.782
PROBING	73	1578	18	0	2	94.434
DOS	27	3	3101	0	0	99.042
U2R	12	2	0	11	15	27.500
R2L	35	1	0	1	441	92.259

表 3　数据量为 7615 条时所提出的 FTBED 算法对各类样本的准确率

实际类别	预测结果/条					准确率/%
	NORMAL	PROBING	DOS	U2R	R2L	
NORMAL	714	24	0	1	33	92.487
PROBING	61	1564	39	1	6	93.596
DOS	29	5	3096	0	1	98.882
U2R	19	0	0	3	18	7.500
R2L	23	30	0	2	423	88.494

表 4 和表 5 分别表示 KNN 算法与所提出的 FTBED 算法在数据量为 27 241 条时对各类样本的准确率.

表 4 数据量为 27 241 条时 KNN 算法对各类样本的准确率

实际类别	预测结果/条					准确率/%
	NORMAL	PROBING	DOS	U2R	R2L	
NORMAL	3922	3	2	0	34	99.015
PROBING	66	1591	17	0	0	95.042
DOS	56	15	15 646	0	1	99.542
U2R	17	0	0	10	11	26.316
R2L	18	0	0	0	431	95.991

表 5 数据量为 27 241 条时所提出的 FTBED 算法对各类样本的准确率

实际类别	预测结果/条					准确率/%
	NORMAL	PROBING	DOS	U2R	R2L	
NORMAL	3884	9	25	0	43	98.056
PROBING	62	1591	21	0	0	95.042
DOS	88	11	15 615	0	4	99.345
U2R	13	0	1	1	23	2.632
R2L	43	1	2	1	402	89.532

表 6 和表 7 分别表示 KNN 算法与所提出的 FTBED 算法在数据量为 51 440 条时对各类样本的准确率.

表 6 数据量为 51 440 条时 KNN 算法对各类样本的准确率

实际类别	预测结果/条					准确率/%
	NORMAL	PROBING	DOS	U2R	R2L	
NORMAL	7740	7	17	0	25	99.371
PROBING	71	1531	19	0	0	94.448
DOS	70	12	31 112	0	3	99.728
U2R	26	0	0	5	11	11.905
R2L	40	3	1	0	394	89.954

表 7 数据量为 51 440 条时所提出的 FTBED 算法对各类样本的准确率

实际类别	预测结果/条					准确率/%
	NORMAL	PROBING	DOS	U2R	R2L	
NORMAL	7706	5	35	1	42	98.934
PROBING	97	1496	26	0	2	92.289
DOS	127	16	31 050	1	3	99.529
U2R	12	0	0	7	23	16.667
R2L	35	6	3	18	376	85.845

实验结果表明,较之 KNN 算法,本文所提出的 FTBED 算法是一种适用于大样本、多特征数据集的算法,尤其在时效性方面,相较于传统的 KNN 算法有着较为明显的优势. 同时,FTBED 算法在准确率方面也有着良好的表现,并且,随着数据量的增加,FTBED 算法与 KNN 算法的准确率在不断接近,时效性的优势也在不断扩大.

4 结束语

本文提出了一种基于可拓距的特征变换方法,该方法首先运用聚类算法得到聚类中心,然后根据聚类的结果将数据点的原特征映射为簇外中心距和簇内可拓距这 2 大部分特征,既可以做到不同类数据的区分,也

能对处于同一类的数据进行区分，最终将 FTBED 应用于网络入侵检测的特征变换中，以达到特征降维的目的，旨在同时满足网络入侵检测系统的实时性和有效性的需求. 相较于传统的 KNN 算法，基于可拓距的特征变换方法明显地减少了检测时间，而同时其检测率的下降可以控制在 1％之内，具有较好的时效性优势.

值得注意的是，网络入侵检测特征变换方法的效率在很大程度上依赖于聚类算法的结果，本文使用 k-means算法聚类时，采取的是随机选择的方法选取初始聚类中心点. 对此，如何较好地选取聚类中心，改进 FTBED 算法的准确率将是下一步工作的核心问题.

参　考　文　献

[1] Tsai C F, Lin C Y. A triangle area based nearest neighbors approach to intrusion detection [J]. Pattern recognition, 2010, 43(1): 222-229.

[2] Tsai C F, Tsai J H, Chou J S. Centroid-based nearest neighbor feature representation for e-government intrusion detection [C]//World Telecommunications Congress. Piscataway: IEEE Press, 2012: 1-6.

[3] 郭春. 基于数据挖掘的网络入侵检测关键技术研究[D]. 北京：北京邮电大学，2014.

[4] Wang X, Zhang C, Zheng K. Intrusion detection algorithm based on density, cluster centers, and nearest neighbors [J]. China Communications, 2016, 13(7): 24-31.

[5] 蔡文. 可拓集和不相容问题[J]. 科学探索学报，1983(1)：83-97.

[6] 李岗. 基于可拓聚类方法的数据挖掘研究[D]. 青岛：中国海洋大学，2009.

[7] 叶永伟，张帆，王运. 基于可拓距的起重机产品配置方法设计[J]. 中国制造业信息化，2012(23)：24-27.

[8] 花黄伟，杨春燕. 可拓识别方法及其在无人车识别障碍物中的应用研究[J]. 广东工业大学学报，2016(04)：1-6.

[9] 朱弘扬，高红，刘巍，等. 可拓 AdaBoost 算法对预测结果的改进[J]. 辽宁工程技术大学学报(自然科学版)，2016(09)：993-997.

[10] 杨春燕，蔡文. 可拓学[M]. 北京：科学出版社，2014.

[11] 王洁松，张小飞. KDDCup99 网络入侵检测数据的分析和预处理[J]. 科技信息(科学教研)，2008(15)：407-408.

[12] 张新有，曾华燊，贾磊. 入侵检测数据集 KDD CUP99 研究[J]. 计算机工程与设计，2010(22)：4809-4812.

A Feature Transformation Method Based on Extension Distance and its Application in Network Intrusion Detection

Xu Hui, Liu Xiang, Fang Ce, Zong Xinlu

(School of Computer Science, Hubei University of Technology, Wuhan 430068, China)

Abstract: As one of the important steps to identify attacks or abnormal behavior to protect network security, network intrusion detection is often used in conjunction with data mining or machine learning techniques. Nowadays, with the explosive growth of network data, the traditional intrusion detection technology is faced with the problem of massive data detection and processing. The existing intrusion detection system is often difficult to meet the real-time demand and the effective demand at the same time. This paper attempts to introduce the concept of extension distance from Extenics into the research of network intrusion detection, and proposes a feature transformation method based on extension distance, which maps the original features of data points into two parts, namely center distance out of the cluster and extension distance in the cluster, the new features are generated according to the multidimensional features of the original data, in order to meet the purpose of reducing feature dimensionality and satisfying the real-time performance and the effectiveness of the network intrusion detection system at the same time. In this paper, KDD CUP 99 data set is used as the simulation data set to test the effectiveness of the proposed method which based on extended distance and using in network intrusion detection. The experimental results show that compared with the traditional KNN algorithm, the new method which based on extended distance can obviously reduce the detection time, and the decrease of the detection rate can be controlled within 1％, so it has a better time advantage.

Keywords: network intrusion detection; feature transformation; Extenics; cluster distance outer center distance; extension distance in the cluster

[责任编辑　杨浦]