基于核的自适应聚类及其在入侵检测中的应用*)

孙 胜[1,2] 王元珍[1]

(华中科技大学计算机学院 武汉 430074)[1] (黄石理工学院计算机学院 黄石 435003)[2]

摘 要 针对k-medoid算法不能有效聚类大数据集和高维数据的弱点,将核学习方法引入到k-medoid算法,提出了基于核的自适应k-medoid算法,使其能够对大数据集和高维数据进行聚类。用KDD 99标准数据集进行实验研究,结果表明该算法性能是优良的,并且能获得令人满意的检测效果。

关键词 聚类,核方法,核函数,k-中心点,异常检测

Kernel-based Adaptive K-medoid Clustering and its Application in Intrusion Detection

SUN Sheng[1,2] WANG Yuan-zhen[1]

(School of Computer Science, Huazhong University of Science and Technology, Wuhan 430074, China)[1]

(School of Computer Science, Huangshi Institute of Technology, Huangshi 435003, China)[2]

Abstract Aiming at the weakness of k-medoid algorithm that it is unable effectively to cluster large data set and high-dimension data, kernel-based learning method was introduced to the k-medoid clustering algorithm, kernel-based adaptive k-medoid algorithm was proposed, and it can cluster large data set and high-dimensional data. The results of experiment using the data sets of KDD cup 99 demonstrate that the algorithm has excellent capability and applying it in intrusion detection system can be an effective way.

Keywords Cluster, Kernel method, Kernel function, K-medoid, Anomaly detection

1 引言

聚类是数据挖掘领域的一项重要研究课题,它是依据样本间关联的度量标准将其自动分为几个群组,且使同一群组内的样本相似,而属于不同群组的样本相异的方法。聚类属于无监督学习,主要体现在聚类学习的数据对象没有类别标记,需要由聚类算法自动计算。聚类方法尤其适合用来探讨样本间的相互关联从而对样本结构做出初步评价,它是从未标记样本中提取出有用信息的过程。一个聚类算法的质量由多个指标来衡量[1],包括算法的可伸缩性、发现任意形状的类的能力、处理高维数据的能力和处理噪声的能力等。

核方法的基本思想是用非线性变换 $\Phi(\cdot)$ 把输入模式空间映射到一个高维特征空间,再在高维特征空间中设计线性学习算法[2]。若算法中各模式矢量间的相互作用仅限于内积运算,则不需要知道非线性变换 $\Phi(\cdot)$ 的具体形式,只要用满足 Mercer 条件的核函数替换线性算法中的内积,就能得到原输入空间中对应的非线性算法。这种称为核技巧的方法,即用 Mercer 核函数来简化计算,可以解决非线性变换带来的维数灾难问题。

本文将核方法引入到k-medoid算法,提出基于核方法的自适应k-medoid算法,使其能够对大数据集和高维数据进行聚类,在聚类过程中,数据可以自适应地加入到最适合它的簇当中。在提出该算法之后,将其应用到基于异常的网络入侵检测系统中,实验结果表明,采用了该算法的入侵检测系统具有较好的检测性能,能够对系统异常情况进行准确预测。

2 基于聚类的异常入侵检测

异常检测在建立正常模型的基础上[3],通过实际行为与之比较是否偏离来判断入侵行为。它能检测已知入侵,也能检测未知入侵。利用聚类算法进行基于异常的入侵检测是建立在以下的两个假设上:(1) 正常行为的数目远远大于入侵行为的数目;(2)入侵的行为和正常的行为差异非常大。该方法的基本思想是由于入侵事件和正常活动事件实质不同并且数目相对很少,因此它们在能够检测到的数据中呈现出比较特殊的特性。

利用核聚类算法的异常检测技术进行检测入侵的思想是:首先将高维数据映射到一个新的高维特征空间中,然后通过特征空间中点的位置来找出离群点,从而实现对入侵的检测[4]。

对数据集进行聚类前,一般要将数据进行标准化处理。标准化数据可以将要考察的数据规范到一个统一的数据超立方体中。数据的标准化处理的方法[5]有很多种,本文采用max-min标准化处理如式(1),该方法可以完整地保存原始数据之间的关系且易于实现。

$$v_i'(j)=\frac{v_i(j)-\min(j)}{\max(j)-\min(j)} \tag{1}$$

其中 $v_i(j)$ 为原连接数据特征值,$\max(j)$,$\min(j)$ 分别为原始连接特征数据中的最大和最小值,$v_i'(j)$ 是规范化后的数据值。

*)基金项目:教育部博士点基金项目(20030487032)。孙 胜 讲师,博士研究生,主要研究方向为现代数据库理论与技术及系统实现;王元珍 教授,博士生导师,主要研究方向为现代数据库理论与技术及系统实现。

3 核方法的相关概念

核学习方法的基本思想是利用核函数 Φ 把输入空间样本映射到定义内积的特征空间 F，使得映射后的样本具有更好的聚类形式。

定义 1 给定集合 D 和核函数 $k:D\times D\to C$，$n\times n$ 阶矩阵 $K=(k(d_i,d_j))_{ij}$ 为集合 D 在核 k 下的 Gram 矩阵(或核矩阵)。若对于任意复数 $c_i\in C$，Gram 矩阵 K_{ij} 满足 $\sum_{ij}c_ic_jK_{ij}\geqslant 0$，则称矩阵 K 为正定的，k 为正定核。

定理 1[6] 对于集合 D 和正定核 k，总存在函数 $\Phi:D\to F$，将集合 D 映射到高维线性空间 F，且满足 $k(d_i,d_j)=(\Phi(d_i)\cdot\Phi(d_j))$，其中 $(\Phi(d_i)\cdot\Phi(d_j))$ 为特征空间 F 中的内积运算。

核函数 $k(\cdot)$[7]计算特征空间 F 中向量之间的相似度，不同的核函数将原始样本空间映射到不同的特征空间。常用的核函数有：

(1)多项式核：$k(x,x_i)=(x\cdot x_i+1)^d$，$d$ 为阶数

(2)RBF 核函数：$k(x,x_i)=\exp(-\frac{\|x-x_i\|^2}{2\sigma^2})$

(3)多层感知机：采用 Sigmoid 函数作为内积，这时就实现了包含一个隐层的多层感知机，隐层节点数目由算法自动确定。满足 Mercer 条件的 Sigmoid 函数为：$k(x,x_i)=\tanh(b(x\cdot x_i)-c)$，式中 b,c 为常数。

(4)B 样条核：$k(x,x_i)=B_{2n+1}(x-x_i)$，其中的核函数需要预先指定。

在特征空间 F，模式向量 x_i 和 x_j 之间的欧氏距离可表示为：

$$d_F(x_i,x_j)=[k(x_i,x_i)-2k(x_i,x_j)+k(x_j,x_j)]^{1/2} \quad (2)$$

给定 n 个对象的数据集合 D，基于核的 k-medoid 算法生成 k 个非空不相交子集 $\{Cl_1,Cl_2,\cdots,Cl_k\}$，使得向量 $\Phi(V_{ij})$ $(1\leqslant i\leqslant k,1\leqslant j\leqslant|Cl_i|)$ 到向量集 Cl_i 的中心点 $\Phi(M_i)$ $(1\leqslant i\leqslant k)$ 的欧氏距离最小。平方差函数为：

$$E_F=\sum_{i=1\cdots k}\sum_{j=1,\cdots,|Cl_i|}\|\phi(M_i)-\phi(x_j)\|^2 \quad (3)$$

4 基于核方法的自适应 k-medoid 聚类算法

基于核的自适应 k-medoid 聚类算法初始从特征空间 F 随机选取 k 个对象，并以这 k 个对象作为起始的中心点。特征空间 F 中任意两个对象之间的欧氏距离用上一节的公式(2)计算，不过在计算距离之前要先规格化对象，使得所有对象都具有相同的属性，并且每种属性在计算过程中权值相等，对象的属性也必须是完备的，没有缺失的属性。

定义两簇中最靠近的两个对象间的距离为簇与簇的距离，对象和簇中最靠近它的对象之间的距离为对象与簇的距离，簇内对象与该簇的距离为零。算法事先计算出任意两簇之间的欧氏距离，它存储在一个二维下三角矩阵当中，还要计算出任意两个最近邻簇之间的距离，同时记录下这两个最近邻簇的簇标号。

对于每个未标记对象 E_i，计算该对象与任意簇之间的距离。假设 E_i 与簇 C_{n1} 的距离为最小，而 C_{n1} 的最近邻簇为 C_{n2}，C_{n1} 和 C_{n2} 的距离为 $d_{\min}$。E_i 的指派可分以下 3 种情形讨论：

(1)如果 E_i 与簇 C_{n1} 的距离为零，则将 E_i 指派给 C_{n1}，然后标记下一个对象；

(2)如果 E_i 与簇 C_{n1} 的距离比 $d_{\min}$ 小，则将 E_i 指派给 C_{n1}，同时更新簇 C_{n1} 的中心点，使得簇内各对象到簇中心的距离之和为最小。另外，在加入 E_i 以后，C_{n1} 与某些簇的距离发生了改变，必须重新计算 C_{n1} 与这些簇的距离，还要重新寻找 C_{n1} 的最近邻簇，并更新 C_{n1} 与其最近邻簇的距离；

(3)如果 E_i 与簇 C_{n1} 的距离比 $d_{\min}$ 大，则合并 C_{n1} 和 C_{n2}，即将 C_{n2} 中的所有对象添加至 C_{n1}，再删除 C_{n2} 内的所有对象，这样 C_{n2} 就变为不含对象的空簇。之后，把 E_i 加入到 C_{n2} 里。同时更新簇 C_{n1} 的中心点，使得簇内各对象到簇中心的距离之和为最小。另外，要重新计算任意两簇的距离，识别出任意两个最近邻的簇以及计算这两个簇的距离。

基于核的自适应 k-medoid 聚类算法的步骤如下：

(1)选择簇的数目 k，并在特征空间 F 中任意选取 k 个样本作为初始的中心点，即初始中心点为 $(M_1,M_2,\cdots,M_k)$；

(2)计算任意两簇的欧氏距离 $d_F(C_i,C_j)=\min_{1\leqslant p\leqslant|C_i|,1\leqslant q\leqslant|C_j|}d_F(x_p,x_q)$；

(3)反复从除了初始中心点之外的其它样本中选取一个未标记对象 E_i，计算该对象与任意簇的距离 $d_F(E_i,C_j)=\min_{1\leqslant p\leqslant|C_j|}d_F(E_i,x_p)$，$j=1,\cdots,k$；

(4)计算 E_i 与其最近邻簇 C_{n1} 的距离 $d_F(E_i,C_{n1})=\min_{1\leqslant j\leqslant k}d_F(E_i,C_j)$；

(5)计算 C_{n1} 与其最近邻簇 C_{n2} 的距离 $d_{\min}=d_F(C_{n1},C_{n2})=\min_{1\leqslant k,j\neq n1}d_F(C_{n1},C_j)$；

(6)根据 $d_F(E_i,C_{n1})$ 与 $d_{\min}$ 的关系，标记对象 E_i，并更新 C_{n1} 与 C_{n2} 的簇成员，同时按照公式(3)重新计算 C_{n1} 的中心点；

(7)重新计算任意两簇的欧氏距离 $d_F(C_i,C_j)$ 和任意两个最近邻的簇的距离；

(8)转至第(3)步，直到标记完所有样本为止。

5 实验结果与分析

实验中采用的样本数据取自 KDD CUP1999 数据，它来源于 1998DARPA 入侵检测评估程序[8]。该数据集大约包含近 500 万个活动事件记录数据对象，每个数据对象均为模拟网络连接过程所发生的 TCP 数据帧。这些数据所涉及的入侵行为共分为 4 大类：拒绝服务攻击(DoS)、未授权的远程访问攻击(R2L)、本地权限提升攻击(U2R)以及扫描与探测攻击(Probe)。

实验过程分为 3 步。首先，对捕捉到的网络数据包进行预处理；第 2 步是将原始数据划分为训练数据和测试数据；第 3 步，使用文中提出的基于核的自适应 k-medoid 聚类算法标记此网络连接记录。

实验环境：应用程序为 Microsoft Visual c++6.0 编写/编译，实验机器为 Intel PⅢ866，256MB 内存。

5.1 数据预处理

实验中采集到的网络数据流有 41 个特征属性，这些特征中很多属于冗余特征，因此只选择其中 16 个能够体现用户行为的部分特征作为研究对象，分别是：duration(连接持续的时间)，protocol_type(协议类型)，service(目标网络服务)，src_bytes(源地址到目标地址的字节数)，dst_bytes(目标地址到源地址的字节数)，flag(网络连接状态，正常或错误)等。

为了满足检测算法中两个假设的需要，从整个检测数据集中提取 5 组数据作为实验用数据，每组数据中包含 130317 条记录，其中 129014 条正常数据，1303 条入侵数据，正常数

(下转第 228 页)

的“数字拓片”图像效果加以比较(注:整个“数字拓片”的制作过程,是严格参照该印泥盒“手工拓片”进行的)。

(2)实验分析:即使是同一实物,由于处在不同光线情况下进行拍摄,所得的图像效果有着明显的差异,将它们按相同方法与步骤进行图像处理后得到对应的“数字拓片”,其图像结果也存在差异。具体表现为:受正面光照射,“实物照片①”中图底(“龙”为图,非“龙”部分为底)的反光比“实物照片②”更加明显,所得到的“数字拓片①”空底部分的斑点就更多。这些斑点对主体(龙)的图像造成一定的干扰,不利于主体图像形成鲜明、直观的视觉效果。相比较而言,实物照片②由于斜顶光的因素,图与底的明暗反差强烈,“龙”形突出,所得“数字拓片②”空底部分的干扰斑点几乎没有,主体图像鲜明。

在良好光线(没有反光)下拍摄得到的“实物照片”,经过photoshop图像处理软件处理后,能够获得与“手工拓片”十分相似的图像效果。

(3)实验初步证明:光线因素对数字拓片的生成影响很大。如果实物表面呈现反光越多,就越不利于生成理想质量的“数字拓片”。因此,若希望生成理想质量的“数字拓片”,就必须使待转换的实物图像尽量满足“没有反光”的拍摄条件。

结束语 本文提出了一种用浮雕实物数字照片来制作“数字拓片”的图像处理方法,它将先进的计算机技术应用到中国传统艺术领域中的探索与思考,不仅能促进艺术领域的发展,而且对计算机技术的拓展有着积极作用。在“数字拓片”技术的研究过程中,为了确保“数字拓片”的可信度,我们采用了样本比对方法;在采集“比对样本”时,充分发挥了艺术专业的学科优势,严格按照传统拓片制作流程及方法,手工拓制了浮雕物的拓片。实验结果表明,采用该方法制作数字拓片的标准化研究方面是行之有效的;但如何实现拓片的风格化和个性化方面,使“数字拓片”增添更丰富的艺术效果将是今后进一步努力的方向。

参 考 文 献

[1] 关文涛. 选择的艺术 Photoshop CS 图像处理深度剖析. 人民邮电出版社,2005

[2] 向世明. Visual C++ 数字图像与图形处理. 北京:电子工业出版社,2002

[3] 李雅梅,吴中福. 基于形态学变换等技术的川南石刻图像预处理方法研究. 计算机科学,2008(3):296-299

[4] Kloskowski M. Photoshop CS2 从入门到精通. 魏海萍,韩滨,于晓菲,译. 北京:电子工业出版社,2006

[5] Braunl T. 并行图像处理. 李俊山,李新社,焦康,译. 西安:西安交通大学出版社,2003

[6] 求是科技. Visual C++数字图像处理典型算法实现. 人民邮电出版社,2006

[7] 蒋先刚. 基于 Delphi 的数字图像处理工程软件设计. 中国水利水电出版社,2006

(上接第 191 页)

据在所有数据中占到了 99%,符合了检测算法中正常数据的数目要远多于入侵数据数目的假设。

5.2 测试结果与分析

入侵检测系统的检测效果主要取决于两个指标,即检测率和误报率。检测率是指被检测出的异常数据样本数占异常样本总数的百分比。误报率则是正常样本被误报为异常的样本数占正常样本数的百分比。实验结果如表 1 所示。

表 1 基于核的自适应聚类算法对入侵的检测情况

样本集	检测率(%)	误报率(%)
第一组测试数据	98.81	0.91
第二组测试数据	99.06	0.75
第三组测试数据	99.37	0.48
第四组测试数据	99.14	0.60
第五组测试数据	99.25	0.72
测试数据总集	99.12	0.69

从表 1 可以看出,系统总检测率达到 99.12%,误报率仅为 0.69%,这表明将基于核的自适应 k-medoid 聚类算法引入到入侵检测中是成功的,系统具有较高的检测率和较低的误报率。

用 k-medoid 聚类算法和基于核的自适应聚类算法对同样的测试数据集进行检测试验,得到图 1 所示的比较曲线。

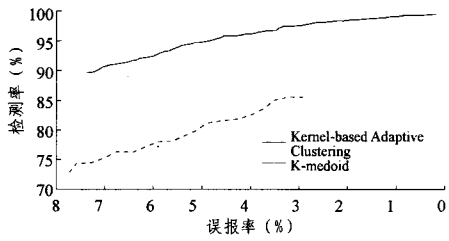

图 1 K-medoid 和基于核的自适应聚类算法入侵检测结果曲线

从图 1 所示的曲线可以反映出,基于核的自适应聚类算法的性能远优于 k-medoid 聚类,其稳定性和可靠性也明显好于 k-medoid 方法。

结束语 本文提出了一种新的高维数据聚类方法,即基于核的自适应 k-medoid 算法,这种算法可以有效地对海量数据集进行聚类。随着网络安全问题的日益突出,入侵检测作为网络安全研究的重要内容,也愈来愈多地受到人们的关注。在提出基于核的自适应 k-medoid 算法之后,将其引入到入侵检测系统中,测试实验结果表明,运用该算法得到了较高的检测率和较低的误报率,同时在与 k-medoid 聚类算法的比较试验中,证实了该算法具有较好的稳定性和可靠性。

参 考 文 献

[1] Fred A L. A new cluster isolation criterion based dissimilarity increments[J]. IEEE Trans. Pattern Analysis and Machine Intelligence,2003,25(6):68-78

[2] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2001,26(1):32-42

[3] 饶鲜,等. 基于支持向量机的入侵检测系统[J]. 软件学报,2003,14(4):798-803

[4] Eskin E, Arnold A, et al. Geometric Framework for Unsupervised Anomaly Detection: Detection Intrusions in Unlabeled Data. Data Mining for Security Applications(DMSA-2002), Kluwer,2002

[5] Han Jiawei, Kamper M. Data Mining: Concepts and techniques [M]. Morgan Kaufmann Publisher,2000:9-10

[6] Scholkopf B. Statistical learning and kernel methods[Z]. MSR-TR2000-23. Microsoft Research,2000

[7] Scholkopf B, Mika S, Burges C, et al. Input space versus feature space in kernel-based method[J]. IEEE Tran. on Neural Networks,1999,10(5):1000-1017

[8] Scholkopf B, Platt J C, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001, 13(7): 1443-1471