ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 2,Mar. 2019,pp. 331－340
DOI:10. 16337/j. 1004-9037. 2019. 02. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于标签传递的异常检测算法研究

赵 曼 赵 耀 朱振峰

(北京交通大学信息科学研究所,北京,100044)

摘 要: 异常检测旨在检测出观测数据中的非正常值,被广泛应用于反信用卡欺诈、网络入侵检测、医疗分析以及气象预报等领域。在异常检测中,正常数据通常具有异常数据所不具备的某种内蕴结构。因此,如何有效地利用正常数据与异常数据在数据结构上的差异性将有助于提高异常检测性能。为此,本文提出了一种新颖的基于标签传递的异常检测算法。该算法通过图模型刻画正常数据所具有的内蕴结构,并通过多重标签传递来构建未标记正例样本与待测试样本的标签置信度的差异。最后,基于正例样本的标签置信度的统计特性分析,实现对测试样本的异常性判决。在人工合成及真实数据集上的实验验证了本文算法的有效性。

关键词: 标签传递;异常检测;图模型

中图分类号: TP391　　**文献标志码:** A

Outlier Detection Based on Label Propagation

Zhao Man, Zhao Yao, Zhu Zhenfeng

(Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China)

Abstract: Outlier detection aims at detecting abnormal values from observational data, which has been used in various files. In outlier detection, normal data are generally embedded in some kind of intrinsic structure that is not suitable for characterizing outliers. Hence, how to effectively utilize the difference in structure between normal data and outliers will contribute to the identification of outlier. Then, a novel label propagation-based outlier detection algorithm is proposed in this paper. To characterize the above intrinsic structure, the graph model is adopted for implementing multiple label propagations. Thus, the difference in structure between normal data and outliers will be identical to the difference of label confidence between them. Furthermore, the statistical characteristic of the label confidences associated to those normal data is explored to give the final ensemble decision on the abnormity of the input test data. The experimental results have validated the effectiveness of the proposed method.

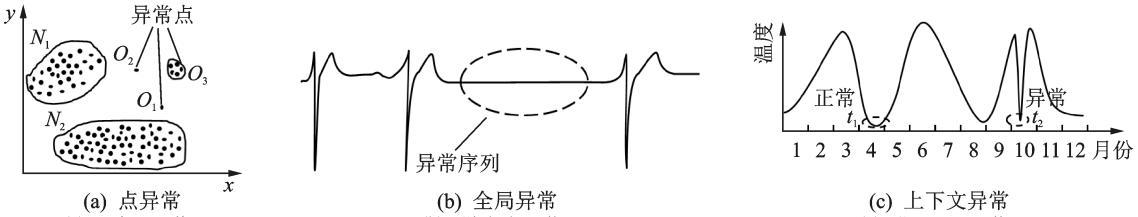
Key words: label propagation; outlier detection; graph model

引 言

随着信息技术的飞速发展,互联网已经渗透到人们日常生活的各个领域。在这些领域中,偶发性的异常常常蕴含着显著的(通常具有很大危害性的)行为信息,如机器的不正常运转表示其存在部件故

基金项目: 国家自然科学基金(61532005, 61572068)资助项目。
收稿日期: 2017-12-02;**修订日期:** 2017-12-26

障,信用卡欺诈意味着巨大的经济损失等[1]。因此,研究如何及时检测异常信息具有重大的现实意义,已成为当前数据挖掘领域中的研究热点[2]。这类检测出观测数据中的非正常值的学习方式,即为异常检测。目前,异常检测主要基于Hawkins对异常的定义[2]:异常点是与其他观测值存在巨大的差异,以至于使人怀疑这些数据产生自不同的机理的观测值。根据异常的定义以及异常和其它数据之间的关系,如图1所示,可以将异常分为3类:点异常、全局异常和上下文异常(条件异常)。



(a) 点异常 (a) Point outlier　(b) 全局异常 (b) Global outlier　(c) 上下文异常 (c) Context outlier

图1 异常类型

Fig.1 Type of outliers

迄今为止学者们提出了多种异常检测算法。根据是否利用数据的标签信息,可以将这些算法分为无监督检测算法和有监督检测算法。

无监督检测算法通过学习样本的隐式结构来区分孤立点或离群点。典型的算法包含局部异常因子(Local outlier factor, LOF)[3-4]、深度算法[5]、基于角度的异常检测算法(Angle-based outlier detection, ABOD)[6]等。由于这类算法没有利用先验标签信息,无法判断其区分出的离群点是否为异常点,此外该类方法对于全局异常和上下文异常也无法有效检测,无监督检测算法并不适用于异常检测,在此不作比较讨论。

有监督检测算法利用已有的标签信息区分异常。在很多场合,异常样本(如卫星网络管理中的故障、网络入侵样例)获取代价极高,数据样本数量严重失衡。基于此,监督检测算法可以分为单分类算法和两分类算法:

(1)单分类算法从正例样本中学习一个数据描述,根据给定或设计的相似性度量准则判定待测样本的类别。典型的算法有基于高斯和小波等的算法[7]支持向量数据描述(Support vector data description, SVDD)[8]及单分类支持向量机(One-class SVM)[9]等基于支撑域的算法、主元分析(PCA)[10]等基于重构的算法以及基于k-mens聚类的检测算法[11]。其中基于K-means聚类的检测算法将正例样本聚为K类,依据每个聚类的中心和半径将不属于所有簇的样本判为异常;SVDD通过建立一个尽可能包含所有正例样本的最小超球检测异常;PCA通过捕捉数据的最大变化方向,利用重构误差区分异常。

(2)由于数据的严重失衡,学者们在现有两分类算法的基础上进行了改进。Veropoulos提出了有偏支持向量机(Biased support vector machine, BSVM)[12],赋予异常较大的惩罚参数;Chawla等人[13]提出了样本合成过采样技术(Synthetic minority oversampling technique, SMOTE)来平衡数据,提高检测性能。

虽然学者们提出了大量的异常检测算法,但目前很少有算法对数据的内蕴结构进行挖掘。而在异常检测问题中,相比于大量存在的正常数据,异常可以被视为一种随机现象,它通常不具有正常数据所具有的某种内蕴结构,因此挖掘数据内蕴结构,利用数据结构差异性有助于提高异常检测性能。为此,本文从数据结构差异性角度出发,提出了基于标签传递的异常检测算法。本文首先标记部分正例样本,通过无向图模型刻画数据所具有的内蕴结构,然后通过多重标签传递获得稳定状态下未标记正例样本和待测数据的标签置信度。由于异常数据不符合正常数据具有的内蕴结构,其在稳态下的标签置信度远远低于正常数据,由此可将它们在数据结构上的差异性转换为标签置信度的差异性。最后基于

正例样本标签置信度的统计特性分析,进行集成判决,实现对测试样本的异常性判决。

1　标签传递模型

1.1　符号说明

为了方便后文阐述,首先对本文用到的一些符号进行说明。令$\boldsymbol{X}=[x_i]_{i=1,2,\cdots,N}\in\mathbf{R}^{N\times d}$为给定的数据集(矩阵),$\boldsymbol{y}=[y_i]_{i\in L}$为标签向量,其中,$y_i\in\{1,2,\cdots,c\}$为与样本$x_i$相对应的类标签,$c$为类别数。此外,定义$\boldsymbol{Y}=[Y_{i,j}]\in\mathbf{R}^{N\times c}$为类标签编码矩阵,其中,$Y_{i,j}=\begin{cases}1 & y_i=j\text{和}i\in L\\ -1 & y_i\neq j\text{和}i\in L\\ 0 & i\in U\end{cases}$,$L=\{1,\cdots,l\}$与$U=\{l+1,\cdots,N\}$分别为标记样本与非标记样本索引集合。基于数据集$\boldsymbol{X}$可构建一个无向图$G(V,E)$,其中$V=\{v_i\}_{i=1,\cdots,N}$为无向图的结点集合,$E=\{e_{i,j}=(v_i,v_j)\}_{i,j=1,\cdots,N}$为无向图的边集合,代表结点间的关系。此外,定义对称权重矩阵$\boldsymbol{W}=[w_{i,j}]_{i,j\in 1,\cdots N}\in\mathbf{R}^{N\times N}$,来反应图结点间的相似度,式中$w_{i,j}=\exp(-\|x_i-x_j\|^2/2\sigma^2)(i\neq j)$,且$w_{i,i}=0$。

1.2　标签传递模型

标签传递(Label propagation, LP)的目的是用已标记数据的标签信息去预测未标记数据的标签信息。Zhou等[14]在高斯场的启发下提出了一种基于图的标签传递模型,该模型本质上属于一种半监督学习模型,其主要思想是:根据样本相似性建立图后,每个样本的标记信息迭代地向其邻近样本传递,直至达到全局稳定状态。

基于全局与局部的一致性假设前提为:(1)临近的样本更可能具有相同的标签;(2)处于同一结构的数据点具有相同标记的可能性较大,构造的标签传递模型可通过最小化如下代价函数得到

$$\min_F Q(F)=\frac{1}{2}\left(\sum_{i,j=1}^{N}w_{i,j}\left\|\frac{F_i}{\sqrt{D_{i,i}}}-\frac{F_j}{\sqrt{D_{j,j}}}\right\|_F^2+\mu\sum_{i=1}^{N}\|F_i-Y_i\|_F^2\right)\tag{1}$$

该式右侧第一项符合一致性假设(1):相邻的数据点具有相似的标记;第二项贴合假设(2),表明样本的稳定状态与其初始标记相关。其中,$\mu>0$为平衡系数,$\boldsymbol{D}=\mathrm{diag}[d_{i,i}]\in\mathbf{R}^{N\times N}$为对角矩阵,$d_{i,i}=\sum_j w_{i,j}$,$F=[F_i,\cdots,F_N]^{\mathrm{T}}\in\mathbf{R}^{N\times c}$为标签置信度矩阵,$\|\cdot\|_F$表示$F$-范数。对于式(1)的最小化问题,不难得出其最优解(标签传递模型)为

$$F=\mu\cdot(\mu\cdot\boldsymbol{I}+\bar{\boldsymbol{\Delta}})^{-1}\cdot\boldsymbol{Y}\tag{2}$$

式中:$\bar{\boldsymbol{\Delta}}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\cdot\boldsymbol{W}\cdot\boldsymbol{D}^{-1/2}$为归一化的拉普拉斯矩阵,$\boldsymbol{I}\in\mathbf{R}^{N\times N}$为单位矩阵。从标签传递的角度,式(2)中的$\left(\mu\boldsymbol{I}+\bar{\boldsymbol{\Delta}}\right)^{-1}$可以看作是一种基于归一化的拉普拉斯矩阵$\bar{\boldsymbol{\Delta}}$形成的扩散核。当直接采用拉普拉斯矩阵$\boldsymbol{\Delta}=\boldsymbol{D}-\boldsymbol{W}$形成扩散核时,由式(2)给出的标签传递模型变为

$$F=\mu\cdot\left(\mu\cdot\boldsymbol{I}+\boldsymbol{\Delta}\right)^{-1}\cdot\boldsymbol{Y}\tag{3}$$

如果把反应原始标签信息的标签编码矩阵$\boldsymbol{Y}$看作标签置信度的初始状态,那么,F则可看作是初始标签置信度经过传递模型后的稳定状态。定理1表明$\boldsymbol{Y}$与F之间存在守恒关系。

定理1　对于式(3)给出的标签传递模型,标签置信度的初始状态与稳定状态之间存在着守恒,即:$1^{\mathrm{T}}\cdot F=1^{\mathrm{T}}\cdot\boldsymbol{Y}$,其中$1\in\mathbf{R}^{N\times 1}$为全1列向量。

证明:用$1^{\mathrm{T}}(\mu\cdot\boldsymbol{I}+\boldsymbol{\Delta})$左乘式(3)两边,可得$1^{\mathrm{T}}\left(\mu\cdot\mathrm{I}+\boldsymbol{\Delta}\right)F=\mu 1^{\mathrm{T}}\cdot\boldsymbol{Y}$。对于拉普拉斯矩阵$\boldsymbol{\Delta}$,由于

存在$\mathbf{1}^{\mathrm{T}}\cdot\boldsymbol{\Delta}=\mathbf{1}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{W})\equiv 0$,为此不难得出$\mathbf{1}^{\mathrm{T}}\cdot\boldsymbol{F}=\mathbf{1}^{\mathrm{T}}\cdot\boldsymbol{Y}$,证明完毕。

基于由式(2),(3)得到的标签置信度矩阵,可通过下式对未标记样本$x_i, i\in U$,的标签置信度进行预测,获得相应的预测标签$\bar{\boldsymbol{y}}_i$,从而实现分类的目的。

$$\bar{\boldsymbol{y}}_i=\mathrm{argmax}_{j\leqslant c}F_{ij} \tag{4}$$

2 基于标签传递的异常检测

2.1 问题描述

对于异常检测问题,相比于大量存在的正常数据,异常数据可以看成是一种随机现象,因而通常不具有正常数据所具有的某种内蕴的数据结构。为此,本文提出一种基于标签传递的异常检测模型。具体来说,即对已知正例样本(训练集)进行部分标记,使这些样本获得初始的标签置信度,然后通过标签传递使得未标记的正常数据以及测试数据获得稳定状态下的标签置信度。对于标记的正例样本,可以假设未标记的正常数据具有与之相一致的数据结构,而异常数据通常不符合该内蕴结构。这样,当通过标签传递达到稳定状态时,未标记的异常样本的标签置信度要远远低于正例样本。因而,可以利用异常样本与正常数据在稳定状态下的标签置信度的差异性,实现对异常样本的有效判决。

图2给出了基于标签传递的异常检测问题的示意图。如图2所示,对于已采集的正常数据,随机标记部分样本,令标记样本的初始标签为1,剩余正常数据与待测数据视作未标记样本,初始标签为0,通过标签传递可获得稳态下各样本的标签置信度。由于异常数据通常不符合正例样本所具有的内蕴结构,这样,异常样本通过标签传递在稳态下的标签置信度明显小于正例样本,因此可用样本标签置信度的差异性体现数据结构上的差异性,进而通过分析稳态下未标记正例样本与待测样本的置信度差异区分异常。

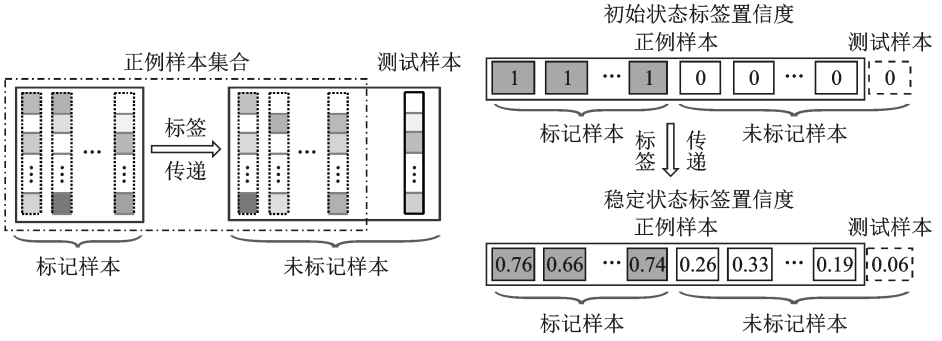


图2 针对异常检测的标签传递示意图

Fig.2 Outlier detection based on label propagation

2.2 异常检测框架

基于2.1节中描述的出发点,本文提出的基于标签传递的异常检测框架如图3所示。令$\boldsymbol{X}\in\mathbf{R}^{N\times d}$表示已有的正例样本数据矩阵,$\boldsymbol{x}_{\mathrm{test}}\in\mathbf{R}^{1\times d}$为待测试数据,$\widetilde{G}(\widetilde{V},\widetilde{E})$为由$\widetilde{\boldsymbol{X}}=[\boldsymbol{X}^{\mathrm{T}}\ \boldsymbol{x}_{\mathrm{test}}^{\mathrm{T}}]^{\mathrm{T}}\in\mathbf{R}^{(N+1)\times d}$构建的无向图。对于来自$\boldsymbol{X}$的$N$个节点,随机选取$l$个样本作为标记样本,其余$N-l$个样本与待测样本共同看作是未标记样本。此时,与标记样本$\widetilde{\boldsymbol{X}}_{\widetilde{L}}$及未标记样本$\widetilde{\boldsymbol{X}}_{\widetilde{U}}$相对应的初始标签分别记为$\widetilde{\boldsymbol{y}}_{\widetilde{L}}=[1]^{\mathrm{T}}\in\mathbf{R}^{l\times 1}$与$\widetilde{\boldsymbol{y}}_{\widetilde{U}}=[0]^{\mathrm{T}}\in\mathbf{R}^{(N+1-l)\times 1}$,其中$\widetilde{L}$与$\widetilde{U}$分别为标记样本与非标记样本的索引集合。经过标签传递后,对于每个未标记样本$\boldsymbol{x}\in\widetilde{\boldsymbol{X}}_{\widetilde{U}}$都将获得一个稳定状态下的标签置信度。进一步,通过对这些标签置信度的统计分析,可以对测试样本进行异常性判决。

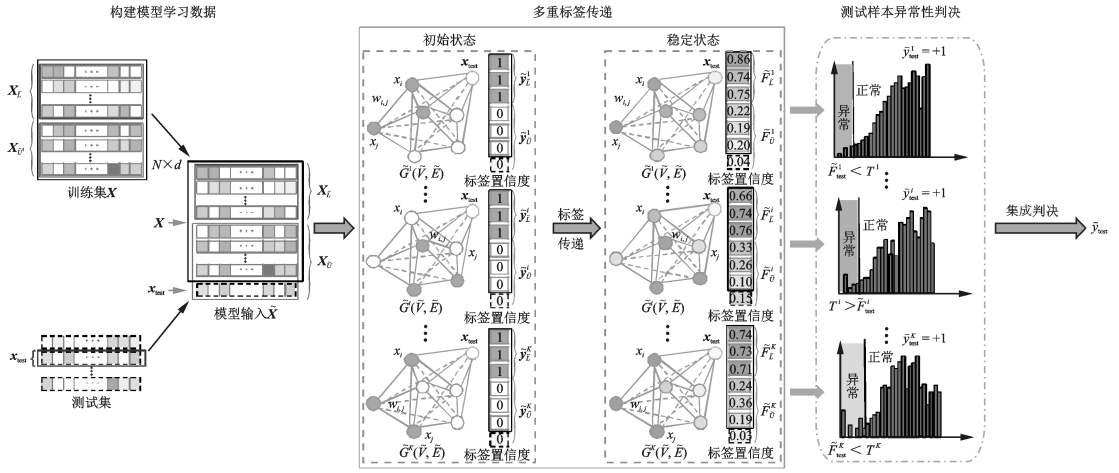


图3　基于标签传递的异常检测框架

Fig.3　Framework of outlier detection based on label propagation

2.3　基于多重随机标记的标签传递

在2.2节中提出,通过从正例样本中随机选取部分样本作为标记样本,然后基于标签传递模型把标记的正例样本的标签置信度传递给未标记的正例样本以及测试样本。这样,可以在稳态下利用异常数据与正例样本标签置信度间的差异性,对测试数据的异常性进行判决。然而,形成上述稳态下标签置信度差异性的前提是假设标记的正例样本与未标记的正例样本具有同一的内蕴数据结构。当对所有正例样本进行如1.2节所述的初始随机标记时,上述假设条件则有时很难满足,进而造成上述差异性不明显。为此,本文提出基于多重随机标记的标签传递,并进一步通过后续的集成判决机制,来克服上述问题。

3　基于标签置信度统计特性的异常判决

对于如图3所示的第i次,$i=1,\cdots,K$,基于随机标记的标签传递,令$P^i(\cdot)$表示此时未标记正例样本在稳态下的标签置信度$\tilde{F}^i_{U^l}$的分布,其中U^l表示未标记样本正例样本的索引集合。对于测试样本在稳态下的标签置信度$\tilde{F}^i_{\text{test}}$,如前所述,当其不服从$P^i(\cdot)$时,则可对其做出异常判决。为此,对于$P^i(\cdot)$,本文采用如下的阈值设置方式

$$T^i=\text{mean}(s^i_1,\cdots,s^i_l) \tag{5}$$
$$S^i=\text{sort}(\tilde{F}^i_{U^l})$$

式中,$S^i=\{s^i_t\}_{t=1,2\cdots,N-l}$,sort$(\cdot)$表示升序函数。

对于测试样本x_{test}在稳态下的标签置信度$\widetilde{F}^i_{\text{test}}$,若$\widetilde{F}^i_{\text{test}}<T^i$,则得与其相应的标签预测值$\bar{y}^i_{\text{test}}=-1$,否则,$\bar{y}^i_{\text{test}}=+1$。

最后,为对测试样本x_{test}的异常性进一步做出可靠性判决,采取如下的集成投票判决方法

$$\bar{y}_{\text{test}}=\text{sign}\left(\sum_{i=1}^{K}\bar{y}^i_{\text{test}}\right) \tag{6}$$

当$\bar{y}_{\text{test}}=-1$时,则可判定x_{test}为异常数据,否则为正常数据。

4 实验结果与分析

4.1 数据集说明及评价标准

4.1.1 数据集

为验证算法的有效性,本文分别在两个人工数据集和5个真实数据集上进行了验证。

两个人工数据集分别为Trefoil-knot数据集、Two moon数据集,分别包含正例样本数:200,100,异常样本数:50,100。

5个真实数据集分别为:(1)USPS手写体数字数据集,包含数字0~9的灰度图像,取其偶数类做实验;(2)UCID图像压缩数据集,包含图像的一次和二次压缩特征,由于图片经过两次压缩可能被篡改或加入了其他信息,二次压缩特征被视作异常类;(3)Arrhythmia数据库,包含正常心率数据和15类异常心率数据;(4)Isolet口语数据集,来源于文献[15],包含150个人对26个英文字母的两次发音信息;(5)Olivetti人脸数据集,包含40个人的人脸图像,取前10个人的人脸信息作为正常类,其余作为异常类。各真实数据集的信息如表1所示。

表1 测试数据统计信息

Tab. 1 Information of real datasets

属性	数据集				
	USPS	UCID	Arrhythmia	Isolet	Olivetti
样本数	9 298	3 200	452	1 560	400
属性	256	1 372	274	617	4 096

4.1.2 评价标准

为评价异常检测的有效性,本文采用Kubat等人提出的G-means评价指标作为异常检测的评价标准,其定义为

$$G\text{-means}=\sqrt{Q_{se}\times Q_{sq}} \tag{7}$$

式中:$Q_{se}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$表示正常类样本准确率,$Q_{sq}=\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$表示异常类样本准确率,TP,TN,FP,FN的定义如表2所示。从定义可以看出,G-means同时兼顾了正常数据类与异常数据类精度的平均,能更客观的反映异常检测算法的检测性能。

表2 异常检测分类混淆矩阵

Tab. 2 Classification confusion matrix in outlier detection

真实类别	判别为正常类	判别为异常类
正常类	Ture positive (TP)	False negative (FN)
异常类	False positive (FP)	Ture negative (TN)

4.1.3 比较算法

为验证本文提出的基于标签传递的异常检测算法的性能(Outlier detection based on label propagation, ODLP),实验中同如下具有代表性的异常检测算法进行了比较:主元分析(PCA)[10]、支持向量数据描述(SVDD)[8]、PCA与SVDD相结合的PSVDD[8,10]以及基于K-means聚类[11]的检测算法。实验中,本文统一令随机标记的正例样本数$l=0.7\times N$,并且进行$K=10$次基于随机标记的标签传递。对于图权重$w_{i,j}=\exp(-\left\|x_i-x_j\right\|^2/2\sigma^2)$中$\sigma$的选取,令$\sigma=k\cdot\mathrm{dist}$,其中dist表示数据集$X$全部样本距离的均值,定义为

$$\mathrm{dist}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left\|x_i-x_j\right\|_F}{N^2} \tag{8}$$

参数k的取值在(0,5)之间选取即可。

4.2　实验结果分析

4.2.1　人工数据集实验结果分析

本文首先在3个人工数据集上进行了验证。图4,5分别展示了针对人工数据集Trefoil-knot和Two moon异常检测示意。图(a)所示为部分正例样本具有标记的初始标签状态,图(b)表示经过标签传递达到稳定状态后的未标记样本标签置信度,置信度越高则颜色越深。从图4,5中可以看出,异常样本的置信度明显小于正例样本,进而可以通过分析稳定状态下未标记样本的置信度进行异常判决。

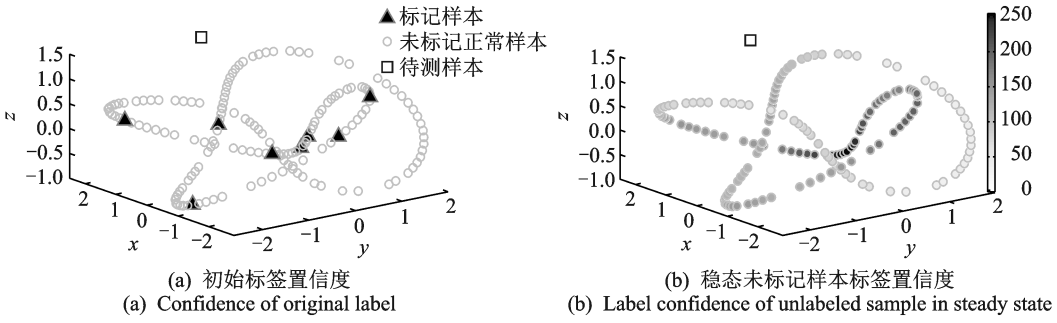


图4　Trefoil-knot数据集实验结果示意图

Fig.4　Schematic diagram of experimental results of trefoil-knot dataset

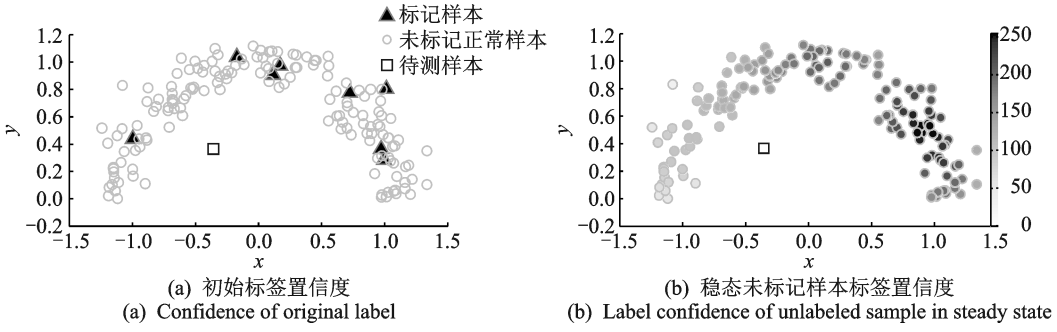


图5　Moon数据集实验结果示意图

Fig.5　Schematic diagram of experimental results of Moon dataset

表3所示为不同算法在人工数据集上的性能对比。可以看出,对于具有流形内蕴结构的合成数据,本文提出的ODLP的检测结果明显优于其他算法。

表3　不同算法在人工数据集上的性能对比

Tab. 3　Performance of different algorithms on artificial datasets　%

数据集	算　法			
	SVDD	PCA	*K*-means	LPOD
Trefoil-knot	96.32±0.85	95.70±0.63	92.19±1.78	99.70±0.49
Moon	97.51±2.08	71.04±2.51	98.43±0.98	99.77±0.48

4.2.2　真实数据集实验结果分析

对于多类数据集USPS,本文采用one-against-all的实验方法,相应的,对于UCID等其他真实数据集,本文统一选取各数据集80%的正例样本作为训练集,剩余样本作为测试集。在实验中,SVDD及

PSVDD算法选用高斯核函数,各种算法的检测结果如表4所示。

表4 不同算法在真实数据集上的性能对比

Tab. 4 Performance of different algorithms on real world datasets %

数据集	算法				
	SVDD	PCA	PSVDD	*K*-means	LPOD
USPS	87.71±1.01	92.40±1.18	82.03±1.73	92.43±0.99	94.10±1.08
UCID	73.69±2.93	94.90±0.86	72.93±0.83	94.03±1.09	96.36±1.36
Arrhythmia	69.38±2.21	68.11±1.72	73.09±1.95	71.15±2.56	73.01±2.48
Isolet	70.72±2.25	80.44±1.61	71.86±2.19	81.52±1.30	83.08±1.67
Olivetti	67.86±2.40	60.43±2.19	71.60±2.50	76.45±2.79	88.16±2.42

从实验结果可以看出,基于*K*-means聚类的算法的检测效果略优于或相近于构建整体数据包络面的SVDD算法和基于重构误差的PCA算法,但对于具有流形结构的数据,其检测效果并不理想。本文提出的ODLP算法由于充分挖掘了数据内蕴结构,将正例样本和异常数据在数据结构上的差异转换为标签置信度的差异,有效利用了数据的局部结构信息,取得了较为理想的检测性能,相较于其他算法,平均提高了2%～3%。

对于流形结构数据集Olivetti,本文提出的ODLP算法能够达到88%的检测性能,明显优于其他算法。

值得注意的是,Arrhythmia数据集相较于其样本维度来说,是一个高维小样本数据集,现有算法均不能取得较理想的检测结果,本文提出的基于标签传递的异常检测算法利用多重随机标记机制成功解决了高维小样本问题,取得了理想的检测效果。

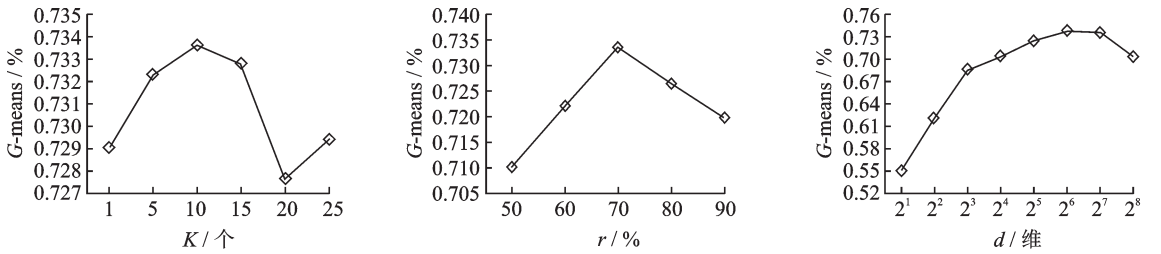
4.2.3 *参数分析*

(1) 集成模型的个数*K*对最终决策结果的影响

图6(a)体现了Arrhythmia真实数据集中集成个数*K*对最终结果的影响,从图中可以看出,当*K*过小或过大时,最终决策效果相应下降,*K*的选取在5～15之间时,检测效果最好,因此本文选取*K*=10作为集成模型的个数。当*K*过小时,由于标记正常数据的随机性,标记的正例样本与未标记的正例样本未必具有同一的内蕴数据结构,因此需要多次随机标记。当*K*过大时,过多的随机标记过程中极有可能出现多次不能反映数据结构的标记。

(2) 标记比例*r*对检测结果的影响

图6(b)所示为随机标记过程中标记比例*r*对检测结果的影响。从图6中可以看出,当标记比例为



(a) 集成个数*K*对结果的影响 (a) Effect of integrated number *K*

(b) 标记比例*r*对结果的影响 (b) Effect of label proportion *r*

(c) 降维维度*d*对结果的影响 (c) Effect of reduced dimension

图6 Arrhythmia数据集中参数的影响

Fig.6 Influence of parameters in Arrhythmia dataset

0.7时,异常检测效果最佳。因标签传递模型的前提是:标记的正例样本与未标记的正例样本具有同一的内蕴数据结构,即正常数据自身的内蕴结构。当标记比例过低时,标记样本不能反映数据的结构;当标记比例过高时,未标记样本数量减少,同样不具有数据的内蕴结构。

(3) PSVDD算法中降维维度对检测结果的影响

PSVDD算法通过PCA降维后使用SVDD模型判决异常,其中降维维度对异常检测结果影响较大,图6(c)所示为Arrhythmia数据集中使用PCA降维后,维度对检测检测结果的影响。通过观察可以发现,原274维数据集降至64维后检测结果较好,因Arrhythmia数据集训练样本较少,数据维度相对较高,所以相应降低数据维度可以提高检测效果,但过低的维度可能导致数据信息丢失严重,影响检测效果。

5　结束语

针对现有异常检测算法很少对数据内蕴结构进行挖掘的缺陷,本文提出了一种基于标签传递的异常检测算法。该算法充分挖掘正常数据与异常数据结构上的差异,首先标记部分正例样本,通过图模型刻画数据的内蕴结构,利用标签传递的思想,提出了标签置信度的概念,并巧妙地将正常数据和异常数据结构上的差异转化为稳定状态下标签置信度的差异;然后通过多重标记过程来避免算法标记样本出现的误差;最后,基于正例样本的标签置信度的统计特性分析,实现对测试样本的异常性判决。在人工合成数据集和真实数据集上的实验,验证了本文算法的有效性。

参 考 文 献:

[1] 薛安荣, 鞠时光, 何伟华, 等. 局部离群点挖掘算法研究[J]. 计算机学报, 2007, 30(8): 1455-1463.
Xue Anrong, Ju Shiguang, He Weihua, et al. Study on algorithms for local outlier detection[J]. Chinese Journal of Computers, 2007, 30(8): 1455-1463.

[2] Larose D T. Discovering knowledge in data: An introduction to data mining[M]. 2nd edition. Hoboken: John Wiley & Sons, 2014: 2: 24.

[3] Radovanović M, Nanopoulos A, Ivanović M. Reverse nearest neighbors in unsupervised distance-based outlier detection[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(5): 1369-1382.

[4] Breunig M M, Kriegel H T, Ng R P, et al. LOF: Identifying density-based local outliers[C]//Proceedings of ACM International Conference on Management of Data (SIGMOD). Dallas, Texas, USA: ACM, 2000: 93-104.

[5] Chen Y, Dang X, Peng H, et al. Outlier detection with kernelized spatial depth function[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2008, 31(2): 288-309.

[6] Kriegel H P, Schubert M S, Zimek A. Angle-based outlier detection in high-dimensional data[C]//Proceedings of the ACM Knowledge Discovery and Data Mining (SIGKDD). Las Vegas, USA: ACM, 2008: 444-452.

[7] 王传旭, 董晨晨. 基于时空特征点的群体异常行为检测算法[J]. 数据采集与处理, 2012, 27(4): 422-428.
Wang Chuanxu, Dong Chenchen. Abnormal crowded behavior detection algorithm based on spatial temporal interesting points [J]. Journal of Data Acquisition & Processing, 2012, 27(4): 422-428.

[8] Tax D M J, Duin R. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.

[9] 潘志松, 陈斌, 缪志敏, 等. One-class分类器研究[J]. 电子学报, 2009, 37(11): 2496-2503.
Pan Zhisong, Chen Bin, Miao Zhimin, et al. Overview of study on one-class classifiers[J]. Chinese Journal of Electronics, 2009, 37(11): 2496-2503.

[10] Ju F, Sun Y, Gao J, et al. Image outlier detection and feature extraction via L1-norm-based 2D probabilistic PCA[J]. IEEE Transactions on Image Processing, 2015, 24(12): 4834-4846.

[11] Marghny M H, Taloba A I. Outlier detection using improved genetic K-means[J]. International Journal of Computer Applications, 2014, 28(11): 620-622.

[12] Veropoulos K, Campbell C, Cristimanini N. Controlling the sensitivity of support vector machines[C]//Proceedings of the International Joint Conferences on Artificial Intelligence(IJCAI). Stockholm, Sweden: Morgan Kaufmann, 1999: 55-60.

[13] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Machine Learning Research, 2002, 16(1): 321-357.

[14] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[C]//Proceedings of the Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia: MIT Press, 2004, 16(16): 321-328.

[15] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples[J]. Journal of Machine Learning Research, 2006, 7(1): 2399-2434.

作者简介:

赵曼(1991-),女,硕士研究生,研究方向:数据挖掘、机器学习,E-mail:14120342@bjtu.edu.cn。

赵耀(1967-),男,教授,研究方向:图像与视频编码、数字水印与信息隐藏、基于内容的信息检索等。

朱振峰(1974-),男,教授,研究方向:图像与视频理解、计算机视觉、机器学习等。

(编辑:张彤)