计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200600170

高阶多视图离群点检测

钟颖宇　陈松灿
南京航空航天大学计算机科学与技术学院　南京 211106
(zhongyingyu@nuaa.edu.cn)

摘　要　由于数据在不同视图之间的分布比较复杂,传统的单视图离群点检测方法不再适用于多视图离群点的检测,使得多视图离群点检测成为一个颇具挑战性的研究课题。多视图离群点可分为3种类型:属性离群点、类离群点和类-属性离群点。现有方法采用跨视图成对约束来学习新的特征表示,并根据这些特征来定义离群点评分度量。这些方法没有充分利用视图间的交互信息,并且在面对3个或更多视图时会导致计算的复杂度更高。为此,文中考虑将多视图数据重塑成张量集形式,定义高阶多视图离群点,并且证明现有的三类多视图离群点都满足高阶多视图离群点的定义,从而提出一种新的多视图离群点检测算法——高阶多视图离群点检测算法(High-Order Multi-View Outlier Detection,HOMVOD)。该算法首先将多视图数据重塑成张量集形式,然后学习其低秩表示,最后设计张量表示下的离群值函数来实现检测。在UCI数据集上的实验表明,HOMVOD算法在检测多视图离群点方面优于现有方法。

关键词:多视图离群点检测;多视图学习;异常检测;张量表示;低秩表示
中图法分类号　TP181

High-order Multi-view Outlier Detection

ZHONG Ying-yu and CHEN Song-can
College of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,Nanjing 211106,China

Abstract　Due to the complex distribution of data between different views,the traditional single-view outlier detection method is no longer applicable to the detection of multi-view outliers,making multi-view outlier detection a challenging research topic. Multi-view outliers can be divided into three types:attribute outliers,class outliers,and class-attribute outliers. Existing methods use pairwise constraints across views to learn new feature representations and define outlier scoring metrics based on these features,which do not take full advantage of the interactive information between views and results in higher computational complexity when facing three or more views. Therefore,this paper considers to reshape multi-view data into tensor set form,defines high-order multi-view outliers,and proves that all of the existing three types of multi-view outliers meet the definition of high-order multi-view outliers,so as to propose a new multi-view outliers detection algorithm called high-order multi-view outliers detection algorithm (HOMVOD). Specifically,the algorithm firstly reshapes multi-view data into tensor set form,then learns its low-rank representation,and finally designs outlier function under tensor representation to realize detection. Experiments on UCI datasets show that this method is superior to existing methods in detecting multi-view outliers.

Keywords　Multi-view outlier detection,Multi-view learning,Anomaly detection,Tensor representation,Low-rank representation

1　引言

离群点检测,也被称为异常检测,是一种用于识别数据集中异常样本的数据分析技术,现已被广泛应用于诸多领域,如入侵和欺诈检测[1-3]、网络垃圾邮件检测[4]。近年来,研究者提出了大量离群点检测方法,如基于分布的方法[5]、基于距离的方法[6-8]、基于密度的方法[9-10]和基于聚类的方法[11-12]。然而,这些离群点检测算法都是针对单视图数据设计的,对于多视图离群点检测场景并不适用。

在现实中,许多数据都来自不同的域或不同的特征提取器,相应的每组特征可被视为一个特定视图,由此形成了多视图数据。近年来,越来越多的研究开始关注此类数据[13-16]。由于特征提取常受到噪声的干扰,使得多视图数据中易出现异常点,进而影响后续任务。因此,研究者开始关注如何从多视图数据中检测出离群点。

最初,多视图离群点检测主要关注的是在每个视图中均表现出异常行为的离群点。这种离群点被定义为属性离群点。然而,属性离群点其实可从单个视图中检测获得,并不需要额外学习多视图的互补信息。之后,有研究提出横向离群点的概念,使得多视图离群点检测开始关注这类跨视图(即横

到稿日期:2020-06-28　返修日期:2020-07-29　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金重点项目(61732006)
This work was supported by the Key Program of National Natural Science Foundation of China (61732006).
通信作者:陈松灿(s.chen@nuaa.edu.cn)

向)特征(如聚类关系)不一致的离群点[17]。此后,大多数多视图离群点检测方法转向关注横向离群点(类离群点)或属性离群点[18-20]。最近,研究者又发现了类-属性离群点问题[21]。如图1所示,黑色三角形表示属性离群点,它是在每个视图中都表现出异常行为的样本;黑色圆圈表示类离群点,对应于横向离群点,它在每个单独的视图中表现正常,但在不同视图中则显示出不一致的特征或行为(如聚类关系);第三类离群点是类-属性离群点,用黑色方块表示,它在某些视图中表现为属性离群点,在其他视图中表现为类离群点。

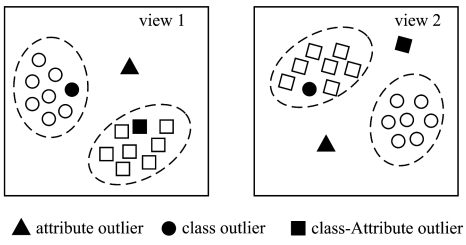


图1 3种离群值的描述

Fig. 1 Illustration of three types of outliers

然而,现有方法都是基于多视图一致性原则,约束视图公共表示或者聚类结果完全一致,通过捕获视图的不一致来检测离群点,并没有充分利用多视图的交互信息。此外,现有方法大多采用视图间成对的方式来识别离群点,从而在面对3个或多个视图时不可避免地会导致较高的复杂度。张量积为张量理论提供了一个优雅的代数结构。这种结构使张量在表示多视图数据时能够充分捕获数据的不同视图副本的关系,同时也能避免视图间的成对比较。因此,文中将多视图数据重塑为张量,提出基于张量表示的多视图离群点检测算法HOMVOD。该算法首先将多视图数据重塑成张量集形式,然后学习其低秩表示,最后设计张量表示下的离群值函数来实现检测。据我们所知,这是基于张量表示的多视图离群点检测算法的首次尝试。本文的主要贡献如下:1)证明了多视图数据的张量表示对多视图离群点检测场景的适用性;2)提出了一种基于低秩表示学习的高阶多视图离群点检测算法;3)UCI数据集上的实验结果验证了本文算法的优越性。

2 相关工作

本节简要回顾了相关的研究工作,主要包括多视图离群点检测和基于张量表示的多视图学习。

2.1 多视图离群点检测

传统的异常点检测方法侧重于在单视图数据集中检测偏离大多数样本的异常样本[1-12]。随着多视图学习的发展[13-16],多视图离群点检测越来越受到重视。最初,多视图离群点检测主要关注的是属性离群点。在提出类离群点(横向离群点)概念之后,许多研究者开始关注类离群点。横向异常检测(Horizontal Anomaly Detection,HOAD)算法是首个尝试检测跨视图特征不一致的离群点的方法[17]。它构造了一个多视图的集成相似矩阵,并通过谱聚类算法学习实例在不同视图的表示来检测离群点。基于亲和传播聚类的多视图异常检测[18](anomaly detection via Affinity Propagation,AP)是关注横向离群点的另一种经典算法。该方法在不同视图中执行亲和传播聚类,并从聚类结果中派生出实例的亲和度向量,最后通过比较不同视图的亲和度向量来检测异常。然而,这两种算法都只能检测横向离群点。为了同时检测属性离群点和类离群点,一些研究通过在低秩子空间学习方法[19]和在k-means聚类方法[20]中使用 $l_{2,1}$ 范数来约束误差矩阵,从而可同时检测属性离群点和类离群点。但是,这些方法都需要使用类标签信息,这使得它们不适用于缺少类标签信息的情形。此外,前述所有方法都必须以视图间成对比较的方式来识别离群值,当面对3个或更多的视图时,会导致计算的复杂度更高。为了解决成对比较的问题,有学者提出另一种基于低秩子空间学习的方法[21]。该方法提出第三类离群点,并通过分离视图公共表示和视图特定表示来克服成对比较的限制。但是,现有方法都是基于多视图一致性原则,约束视图公共表示或者聚类结果完全一致,通过捕获视图的不一致来检测离群点,并没有充分利用多视图的交互信息。

2.2 基于张量表示的多视图学习

假设每个视图都对应一个张量模式,那么多视图数据可以自然地重塑成张量集的形式,相比于传统的表示形式,这样更有利于充分捕获多视图的潜在交互作用[22]。近年来,开展了大量采用张量表示进行多视图学习的研究工作。多视图数据的张量典型相关分析[22]将多视图数据表示为张量,并定义了多视图协方差张量,以此将典型相关分析扩展到多视图场景,从而避免了视图间的两两比较。考虑到传统的多视图特征选择方法没有充分利用视图间的内在相关性,有学者提出一种基于张量的多视图特征选择对偶方法[23]。也有学者将多视图数据的张量表示形式用于多视图聚类任务,提出一种基于多视图数据张量表示的多视图聚类框架[24]。以往的研究表明,多视图张量表示在聚类、相关分析等方面都表现出良好的效果,说明它满足多视图的一致性准则。然而,多视图张量表示是否适用于多视图离群点检测场景,至今尚未有相关的研究。

3 多视图张量表示

本节将分析和证明多视图数据的张量表示形式在多视图离群点检测场景中的适用性。

用 $D=\{X^1,X^2,\cdots,X^M\}$ 表示 M 个视图数据的集合,其中 $X^v\in R^{d_v\times N}$ 表示第 v 个视图中的 N 个样本,样本维度为 d_v。

定义1(多视图张量) 对于每个实例 i,可以构造一个多视图高阶张量 T_i,即 $T_i=x_i^1\circ x_i^2\circ\cdots\circ x_i^v\circ\cdots\circ x_i^M$,其中 $x_i^v\in R^{d_v\times 1}$ 表示第 v 个视图的第 i 个样本,$\circ$ 表示张量积。

下面将解释并证明多视图张量表示满足多视图的一致性准则。

定理1(一致性) 假设多视图数据集 D 满足多视图一致性,考虑两个实例 i 和 j,如果在任何第 v 个视图上都满足 $\|x_i^v-x_j^v\|_F^2\leqslant\sigma(\sigma\to 0)$,那么它们的多视图张量将满足 $\|T_i-T_j\|_F^2\leqslant\sigma(\sigma\to 0)$。

证明:假设在第 v 个视图中,数据可以由同一簇中的数据表示,那么可以得到 $\|x_i^v-x_j^v\|_F^2=\|X^vz_i^v-X^vz_j^v\|_F^2\leqslant\sigma(\sigma\to 0,i\neq j)$。由此可以定义 $z_i^v\approx z_j^v=z^v$。考虑到视图一致性,可以假设 $z^v=z^*$ $(v=1,2,3,\cdots,M)$,得到 $T_i\approx T_j=\circ_{v=1}^M(X^vz^*)$,$\|T_i-T_j\|_F^2\leqslant\sigma(\sigma\to 0)$ 得证。

在证明多视图张量表示的一致性之后,给出高阶多视图离群点的定义,并就此类离群值是否适用于多视图离群点检测进行讨论。

定义 2(高阶多视图离群点) 高阶多视图离群值是指在多视图张量集中远离任何其他正常样本的数据点。

定理 2 如果一个实例属于多视图离群点(属性离群点、类离群点和类-属性离群点),那么它的多视图张量一定是高阶多视图离群点。

证明:

(1)假设实例 i 是正常的,实例 j 是属性离群点,那么在每个视图上都可以得到:

$$\| x_i^v - x_j^v \|_F^2 = \| X^v(z_i^v - z_j^v) \|_F^2 = \| X^v(z_i - z_j^v) \|_F^2 > \sigma(\sigma \to 0, i \neq j)$$

定义 $z_j^v = z_i + \varepsilon_j^v(\varepsilon_j^v \gg 0)$,有:

$$T_i = (\circ_{v=1}^M X^v) z_i$$

$$T_j = (\circ_{v=1}^M X^v(z_i + \varepsilon_j^v)) = (\circ_{v=1}^M X^v) z_i + \cdots + (\circ_{v=1}^M X^v)\varepsilon_j^v$$

由此可得,$\| T_i - T_j \|_F^2 > \sigma(\sigma \to 0)$,那么实例 j 的多视图张量是高阶多视图离群点。

(2)假设实例 i 是正常的,实例 k 是类属性离群点,那么我们假设实例 k 只在第 m 个视图中表现为离群点,在其他视图中则表现正常:

$$\| x_i^m - x_k^m \|_F^2 = \| X^m(z_i^m - z_k^m) \|_F^2 > \sigma(\sigma \to 0)$$

$$\| x_i^v - x_k^v \|_F^2 \leqslant \sigma(\sigma \to 0; v \neq m, v = 1, 2, 3, \cdots, M)$$

根据定义 1,我们有 $T_i = (\circ_{v=1}^M X^v) z_i$。

定义 $z_k^m = z_i + \varepsilon_k^m(\varepsilon_k^m \gg 0)$,可得:

$$T_k = \circ_{v=1, v\neq m}^M (X^v z_k^v) \circ (X^m z_k^m) = \circ_{v=1, v\neq m}^M (X^v z_i) \circ (X^m(z_i + \varepsilon_k^m)) = (\circ_{v=1}^M X^v) z_i + (\circ_{v=1, v\neq m}^M X^v) z_i \circ (X^m \varepsilon_k^m)$$

因此,$\| T_i - T_k \|_F^2 > \sigma(\sigma \to 0)$,实例 k 的多视图张量是高阶多视图离群点。

(3)假设实例 i 和 j 是正常的(在所有视图中 i 和 j 属于不同聚类),实例 k 属于类离群点,那么我们假设实例 k 只在第 m 个视图中与 j 同簇,在其他视图中与 i 同簇,则有:

$$\| x_i^v - x_j^v \|_F^2 = \| X^v(z_i^v - z_j^v) \|_F^2 > \sigma(\sigma \to 0; i \neq j) \Rightarrow z_i^v = z_i \neq z_j^v = z_j$$

$$\| x_i^v - x_k^v \|_F^2 = \| X^v(z_i - z_k^v) \|_F^2 < \sigma(\sigma \to 0) \Rightarrow z_k^v = z_i$$

$$\| x_j^m - x_k^m \|_F^2 = \| X^m(z_j - z_k^m) \|_F^2 < \sigma(\sigma \to 0; v \neq m) \Rightarrow z_k^m = z_j$$

由于 $z_i \neq z_j$,我们定义 $z_j = z_i + \varepsilon_k(\varepsilon_k \gg 0)$,有:

$$T_i = (\circ_{v=1}^M X^v) z_i, T_j = (\circ_{v=1}^M X^v) z_j$$

可得:

$$T_k = (\circ_{v=1}^M X^v) z_i + (\circ_{v=1, v\neq m}^M X^v) z_i \circ (X^m \varepsilon_k) \Rightarrow \| T_i - T_k \|_F^2 > \sigma(\sigma \to 0)$$

$$T_k = (\circ_{v=1}^M X^v) z_j - (\circ_{v=1, v\neq m}^M X^v) z_j \circ (X^m \varepsilon_k) \Rightarrow \| T_j - T_k \|_F^2 > \sigma(\sigma \to 0)$$

由此可得实例 k 的多视图张量是高阶多视图离群点。

由以上分析可知,所有类型的多视图离群点(属性离群点、类离群点和类-属性离群点)都满足高阶多视图离群点的定义,因此多视图离群点检测可转化为高阶多视图离群点检测,故基于聚类的传统单视图离群点检测框架也适用于多视图离群点检测。

4 高阶多视图离群点检测

本节将描述 HOMVOD。本文的目标是通过检测高阶多视图离群点来解决多视图离群点的检测问题。HOMVOD 算法包含 3 个步骤,分别是多视图张量构造、低秩表示学习和离群值的计算。

4.1 多视图张量构造

用 $D = \{X^1, X^2, \cdots, X^M\}$ 表示 M 个视图数据的集合,其中 $X^v \in R^{d_v \times N}$ 表示第 v 个视图中的 N 个样本,特征维度为 d_v。按照上一节构造多视图张量的方式,可以得到:

$$X_i = x_i^1 \circ x_i^2 \circ \cdots \circ x_i^v \circ \cdots \circ x_i^M \in R^{d_1 \times d_2 \times \cdots \times d_M} \tag{1}$$

由此可以把 N 个实例的多视图数据集构造成一个多视图张量集 $I = \{X_i\}_{i=1}^N$,X_i 表示第 i 个实例的多视图张量。

4.2 低秩表示学习

首先,把每个张量 X 展开成向量的形式 $\boldsymbol{t} \in R^{d_1 d_2 \cdots d_M \times 1}$,则多视图张量集 I 就转化成矩阵 $\boldsymbol{T} = [t_1 \quad t_2 \quad \cdots \quad t_N] \in R^{d_1 \times d_2 \times \cdots \times d_M \times N}$。

4.2.1 建模

数据通常位于底层的低维子空间中,而不是均匀分布在整个空间中[25]。因此,这些数据点可以用一个低维子空间来表示,具体可表示为:

$$\boldsymbol{T} = \boldsymbol{TZ} + \boldsymbol{E} \tag{2}$$

其中,$\boldsymbol{Z} = [z_1 \quad z_2 \quad \cdots \quad z_N] \in R^{N \times N}$ 为子空间表示矩阵;每个 $z_i \in R^{N \times 1}$ 都是向量的子空间表示;$\boldsymbol{E} \in R^{d_1 \times d_2 \times \cdots \times d_M \times N}$ 是误差矩阵。

假设每个视图中属于同一簇的实例的多视图张量的向量形式在同一簇中,那么同一簇中的向量形式可以从同一子空间中得到,则 $\boldsymbol{Z}$ 应为一个低秩的系数矩阵。因此可以通过求解以下问题来学习多视图数据的压缩系数表示:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}} \| \boldsymbol{Z} \|_* + \alpha \| \boldsymbol{E} \|_{2,1} \quad \text{s.t } \boldsymbol{T} = \boldsymbol{TZ} + \boldsymbol{E} \tag{3}$$

其中,$\| \cdot \|_*$ 表示迹范数,$\| \cdot \|_{2,1}$ 代表 $l_{2,1}$ 范数。

4.2.2 优化

为了求解式(3),首先应将其转化为如下等价问题:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{J}} \| \boldsymbol{J} \|_* + \alpha \| \boldsymbol{E} \|_{2,1} + \mathrm{tr}[\boldsymbol{Y}_1^{\mathrm{T}}(\boldsymbol{T} - \boldsymbol{TZ} - \boldsymbol{E})] \quad \text{s. t. } \boldsymbol{T} = \boldsymbol{TZ} + \boldsymbol{E}, \boldsymbol{Z} = \boldsymbol{J} \tag{4}$$

式(4)可以通过求解以下增广拉格朗日乘子(Augmented Lagrange Multiplier,ALM)问题来解决:

$$\min_{\boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{J}} \| \boldsymbol{J} \|_* + \alpha \| \boldsymbol{E} \|_{2,1} + \mathrm{tr}[\boldsymbol{Y}_1^{\mathrm{T}}(\boldsymbol{T} - \boldsymbol{TZ} - \boldsymbol{E})] + \mathrm{tr}[\boldsymbol{Y}_2^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{J})] + \mu(\| \boldsymbol{T} - \boldsymbol{TZ} - \boldsymbol{E} \|_F^2 + \| \boldsymbol{Z} - \boldsymbol{J} \|_F^2)/2 \tag{5}$$

其中,$\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$ 是拉格朗日乘子,$\mu > 0$ 是惩罚参数。

式(5)中的变量可以用非精确 ALM 算法求解[26],分步优化过程如下:

更新 $\boldsymbol{J}$:通过只保留与 $\boldsymbol{J}$ 相关的项,得到:

$$\boldsymbol{J} = \arg\min \| \boldsymbol{J} \|_*/\mu + \| \boldsymbol{J} - (\boldsymbol{Z} + \boldsymbol{Y}_2/\mu) \|_F^2/2 \tag{6}$$

利用奇异值阈值(Singular Value Thresholding,SVT)算法[27]可以得到该问题的最优解。

更新 $\boldsymbol{Z}$:通过只保留与 $\boldsymbol{Z}$ 相关的项,得到:

$$\boldsymbol{Z} = \arg\min \mathrm{tr}[\boldsymbol{Y}_1^{\mathrm{T}}(\boldsymbol{T} - \boldsymbol{TZ} - \boldsymbol{E})] + \mathrm{tr}[\boldsymbol{Y}_2^{\mathrm{T}}(\boldsymbol{Z} - \boldsymbol{J})] + \mu(\| \boldsymbol{T} - \boldsymbol{TZ} - \boldsymbol{E} \|_F^2 + \| \boldsymbol{Z} - \boldsymbol{J} \|_F^2)/2 \tag{7}$$

令式(7)关于 $\boldsymbol{Z}$ 的导数为 0,则有:

$$\boldsymbol{Z} = (\boldsymbol{I} + \boldsymbol{T}^{\mathrm{T}}\boldsymbol{T})^{-1}(\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T} - \boldsymbol{T}^{\mathrm{T}}\boldsymbol{E} + \boldsymbol{J} + (\boldsymbol{T}^{\mathrm{T}}Y_1 - Y_2)/\mu) \tag{8}$$

更新 $\boldsymbol{E}$:通过只保留与 $\boldsymbol{E}$ 相关的项,得到:

$$\boldsymbol{E}=\arg\min \alpha \| \boldsymbol{E} \|_{2,1}/\mu+\| \boldsymbol{E}-(\boldsymbol{T}-\boldsymbol{TZ}+\boldsymbol{Y}_1/\mu) \|_F^2/2 \tag{9}$$

已有研究对式(9)的求解进行了探讨[28]。具体来说,令 $\Omega=\boldsymbol{T}-\boldsymbol{TZ}+\boldsymbol{Y}_1/\mu$,解 E^* 的形式如下:

$$E^*(:,i)=\begin{cases}(\|\Omega(:,i)\|-\alpha)\Omega(:,i)/\|\Omega(:,i)\|, & \text{if } \alpha<\|\Omega(:,i)\| \\ 0, & \text{otherwise}\end{cases} \tag{10}$$

更新拉格朗日乘子:

$$\boldsymbol{Y}_1=\boldsymbol{Y}_1+\mu(\boldsymbol{T}-\boldsymbol{TZ}-\boldsymbol{E}) \tag{11}$$

$$\boldsymbol{Y}_2=\boldsymbol{Y}_2+\mu(\boldsymbol{Z}-\boldsymbol{J}) \tag{12}$$

算法 1 中列出了完整的优化过程。

算法 1 用不精确 ALM 解决问题(5)

输入:数据矩阵 $\mathbf{T}$,参数 α

输出:$\mathbf{Z}$,$\mathbf{E}$

1. 初始化:$\mathbf{Z}=\mathbf{J}=0$,$\mathbf{E}=0$,$\mathbf{Y}_1=0$,$\mathbf{Y}_2=0$,$\mu=10^{-6}$,$\mu_{\max}=10^{10}$,$\rho=1.1$,$\varepsilon=10^{-8}$

2. while not coverage do

2.1. 固定其他变量,按式(6)更新 $\mathbf{J}$

2.2. 固定其他变量,按式(8)更新 $\mathbf{Z}$

2.3. 固定其他变量,按式(10)更新 $\mathbf{E}$

2.4. 固定其他变量,按式(11)和式(12)更新拉格朗日乘子

2.5. 按 $\mu=\min(\rho\mu,\mu_{\max})$更新 μ

2.6. 检查收敛条件 $\|\mathbf{T}-\mathbf{TZ}-\mathbf{E}\|_\infty<\varepsilon$ 和 $\|\mathbf{Z}-\mathbf{J}\|_\infty<\varepsilon$

End while

4.2.3 复杂度分析

本节研究算法 1 的时间复杂性,其最耗时的部分有 $\boldsymbol{J}$ 的低秩优化和求解 $\boldsymbol{Z}$ 的过程中的矩阵乘法以及矩阵求逆。具体来说,$\boldsymbol{J}\in R^{n\times n}$ 的低秩优化耗时 $O(n^3)$。当 n 非常大时,$\boldsymbol{J}$ 的优化相当耗时。幸运的是,根据相关研究[29],$\boldsymbol{J}$ 的 SVD 可以加速到 $O(rn^2)$,其中 r 是 $\boldsymbol{J}$ 的秩。求解 $\boldsymbol{Z}$ 涉及的矩阵乘法和矩阵求逆均耗时 $O(n^3)$。因此,算法 1 的时间复杂度为 $O(n^3)$。降低算法复杂性将是后续工作。

4.3 离群值计算

由定理 2 可知,如果一个实例是一个多视图离群点,那么它的多视图张量一定是高阶多视图离群点,它的向量形式对其他正常实例的向量形式的表示贡献较小。基于优化得到的解 $\boldsymbol{Z}$ 和 $\boldsymbol{E}$,设计如下准则度量样本离群值:

$$o(i)=-\|\boldsymbol{Z}(:,i)\|_F^2+\beta\|\boldsymbol{E}(:,i)\|_F^2 \tag{13}$$

其中,$o(i)$表示第 i 个实例的离群值,$\beta>0$ 是权衡参数。

从数据表示的角度看,实例主要由来自同一簇的实例表示,这也适用于多视图张量之间的表示。例如,如果实例 i 是一个正常实例,则系数 $\boldsymbol{Z}(:,i)$的平方误差会很大,而误差向量 $\boldsymbol{E}(:,i)$会很小,使得离群值也很小。反之,如果实例 i 是一个异常实例,则离群值会比较大。在计算出实例的离群值得分后,若得分 $o(i)$大于阈值,则将实例 i 标记为离群点。

算法 2 给出了 HOMVOD 算法的流程。

算法 2 HOMVOD

输入:多视图数据集 D,阈值 γ

输出:二元离群标签向量 $\mathbf{L}$

1. 归一化每个 x_i^v,$x_i^v=x_i^v/\|x_i^v\|$

2. 构造多视图张量集 I,并将它展开成 $\mathbf{T}$

3. 用算法 1 解决问题(5),得到 $\mathbf{Z}$,$\mathbf{E}$

4. 用式(13)为每个实例计算离群值

5. 生成二元离群标签向量 $\mathbf{L}$.

$$\begin{cases}L(i)=1, & \text{if } o(i)>\gamma \\ L(i)=0, & \text{otherwise}\end{cases}$$

5 实验

5.1 数据集和预处理

实验选用 5 个 UCI 常用数据集 zoo,wine,wdbc,iris 和 letter,其信息如表 1 所列。它们都是没有离群值的多类数据集。需要特别指出的是,letter 数据集中包含 26 个字母的 20000 个样本,每个字母包含 700～900 个样本。为节省评估时间,根据相关研究工作的策略[19],为每个字母随机选择 30 个样本,生成 780 个样本子集用于实验。

表 1 5 个 UCI 数据集的统计信息

Table 1 Statistics of 5 UCI data sets

	zoo	wine	wdbc	iris	letter
instance	101	178	569	150	20000
feature	16	12	30	4	16
class	7	3	2	3	26

按照相关研究工作的方法[19-21],在一定比例下生成含有 3 类离群点的多视图数据。首先将每个样本的特征分解成 V 个特征子集,每个子集对应一个视图。例如,为了生成两个视图数据,可以切割 D 维样本,将前$\lfloor D/2 \rfloor$个维度的特征作为第一个视图,其他维度作为第二个视图。接下来生成离群点。对于属性离群点,随机选择样本,并用随机值替换所有视图中的原始特征。对于类离群点,随机抽取样本对,使其在$\lfloor V/2 \rfloor$视图中交换特征向量,而其他视图保持不变。对于类-属性离群点,随机选择样本对,使其在$\lfloor V/2 \rfloor$视图中交换特征向量,在其他视图中用随机值替换原始特征值。

5.2 基准方法

将 HOMVOD 算法与 4 种经典的多视图离群点检测方法进行比较,具体为:1)HOAD[17];2)AP[18];3)离群点检测的多视图低秩分析(Multi-View Low-Rank Analysis, MLRA)[19];4)用于多视图离群点检测的潜在鉴别子空间表示(Latent Discriminant Subspace Representations, LDSR)[21]。我们做了一些调整,以便这些算法能够在选择的数据集上运行。首先,AP 和 MLRA 算法均受到成对检测方式的限制,没有明确的方法能够将它们扩展到 3 个或更多视图。因此,我们在每对视图中计算它们的离群值,然后取平均值作为最终的结果。对于 HOAD 和 AP,使用欧氏距离作为相似性度量,并报告它们的性能。

5.3 评估设置

为了减少数据生成所产生的偏差对实验的影响,对 5 个 UCI 数据集分别生成 20 个离群点数据集。对于每个数据集,每种方法都在生成的 20 个离群点数据集上进行评估,并报告评估结果的平均值和标准差。

根据文献[17],我们采用受试者工作特征(Receiver Characteristic Operator, ROC)曲线下的面积(Area Under the Curve, AUC)作为评价指标,它代表了真阳性率(True Positive Rate, TPR)和假阳性率(False Positive Rate, FPR)之间的权衡。ROC 曲线的 FPR 和 TPR 定义如下:

$$\begin{aligned}FPR &= FP/(FP+TN) \\ TPR &= TP/(TP+FN)\end{aligned} \tag{14}$$

其中,FP,TN,FN,TP 分别代表假阳性、真阴性、假阴性和真阳性。

5.4 对比实验结果

表 2 是在 5 个数据集上得到的 *AUC* 值（平均值±标准差），每种类型的离群点比率均为 5%，最优结果用粗体显示，次优结果用下划线表示。两视图案例和三视图案例分别为原始数据特征的两划分和三划分。三视图数据的 AP 和 MLRA 的 *AUC* 值是通过对所有三对视图的 *AUC* 值取平均而得到的。

表 2 两视图案例和三视图案例在 UCI 数据集上的 *AUC* 值（均值±标准差）

Table 2 *AUC* values (mean ± standard deviation) on UCI datasets of two views and three views

	zoo		wine		wdbc		iris		letter	
	$v=2$	$v=3$	$v=2$	$v=3$	$v=2$	$v=3$	$v=2$	$v=3$	$v=2$	$v=3$
HOAD	0.55±0.06	0.54±0.06	0.68±0.05	0.50±0.06	0.73±0.08	0.51±0.02	0.72±0.05	0.47±0.05	0.58±0.03	0.51±0.03
AP	0.52±0.07	0.44±0.04	0.80±0.04	0.81±0.02	0.97±0.01	0.92±0.01	**0.90±0.05**	0.67±0.04	0.67±0.03	0.69±0.02
MLRA	0.63±0.09	0.58±0.06	0.68±0.06	0.56±0.06	0.60±0.03	0.51±0.02	0.60±0.09	0.54±0.08	0.52±0.02	0.54±0.02
LDSR	**0.90±0.03**	0.86±0.04	**0.88±0.04**	0.87±0.03	0.93±0.02	0.94±0.02	0.79±0.06	0.74±0.06	**0.85±0.02**	0.82±0.01
HOMVOD	0.89±0.03	**0.90±0.04**	0.87±0.03	**0.90±0.03**	**0.98±0.01**	**0.97±0.02**	0.87±0.05	**0.76±0.05**	0.83±0.02	**0.84±0.02**

从表 2 可以看出，HOMVOD 算法在大多数数据集上的性能优于其他算法。虽然该方法在某些数据集上表现出次优的性能，但仍然非常接近最优结果，而且 HOMVOD 算法在三视图情况下一直保持最优的性能。这是由于大多数对比算法都是采用成对学习策略，使得在 3 个或更多视图的情况下无法充分学习多视图的交互信息。

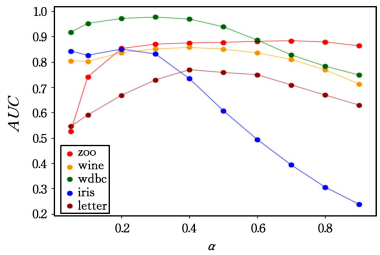
比较两视图和三视图情况下的 *AUC* 值可以看出，大多数方法的性能都有所下降。原因如下：1）在三视图情况下，每个视图的特征维数小于在两视图情况下的特征维数，导致数据聚类结构模糊，离群值结果较差；2）三视图中异常特征与总体特征之比低于两视图。

5.5 分析实验

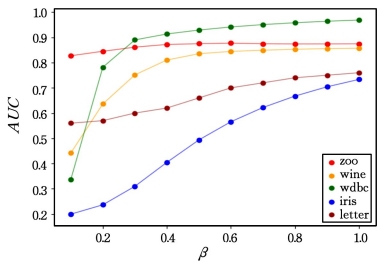
5.5.1 参数分析

HOMVOD 算法有 α 和 β 两个参数，其中 α 用于平衡目标函数，β 用于离群分数的加权。

首先，固定 $\beta=1$，评估 α 的影响。图 2(a)显示了 α 取值为 $\{0.05,0.1,0.2,0.3,0.4,0.5,\cdots,1.0\}$ 时本文方法在 5 个数据集上得到的 *AUC* 值。当 $\alpha\in[0.3,0.5]$ 时，HOMVOD 在大多数数据集上保持了最优的性能。因此，我们将 $\alpha=0.4$ 设置为默认值。



(a)不同 α 取值时本文方法在 5 个数据集上的 *ACU* 值



(b)不同 β 取值时本文方法在 5 个数据集上的 *ACU* 值

图 2 参数分析实验的结果

Fig. 2 Results of parametric analysis experiment

接下来进一步评估固定 $\alpha=0.4$ 后 β 的变化对性能的影响。图 2(b)给出了 β 取值为 $\{0.1,0.2,0.3,0.4,0.5,\cdots,1\}$ 时本文方法在 5 个数据集上得到的 *AUC* 值。可以看出，随着 β 的增大，HOMVOD 在所有数据集上的 *AUC* 值逐渐提升，最后保持在较高值。因此，我们选择 $\beta=1$ 作为默认值。

5.5.2 收敛性分析

本文的计算模型在每次迭代中的相对误差表示为 $\|\boldsymbol{T}-\boldsymbol{TZ}-\boldsymbol{E}\|_F/\|\boldsymbol{T}\|_F$。图 3 给出了 zoo 数据集上模型相对误差随着迭代次数的变化。可以看出，相对误差先快速减小，然后保持稳定。这表明本文模型具有良好的收敛性。

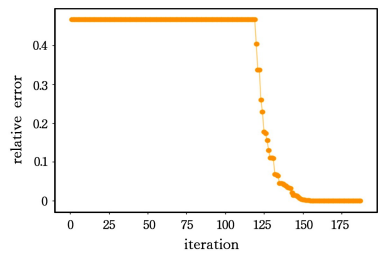


图 3 HOMVOD 的收敛曲线

Fig. 3 Convergence curve of HOMVOD

结束语 为了避免成对学习策略导致的信息丢失和较高的复杂度，充分利用视图间的交互作用，本文提出一种多视图离群点检测算法 HOMVOD。通过将多视图数据重塑成张量集的形式，学习张量集的低秩表示，并设计张量表示下的离群值函数来实现离群点检测。此外，我们对多视图数据的张量表示方式进行了分析，并且定义了高阶多视图离群点，同时证明了传统的 3 种多视图离群点都满足该类离群点的定义，从而可把多视图离群点检测转化为高阶多视图离群点检测，使得传统的单视图离群点检测框架也适用于多视图离群点检测。最后，通过在 UCI 数据集上进行实验验证了所提算法的有效性。

参 考 文 献

[1] WEST J, BHATTACHARYA M. Intelligent financial fraud detection: a comprehensive review [J]. Computers & Security, 2016, 57: 47-66.

[2] BAHNSEN A C, AOUADA D, STOJANOVIC A, et al. Feature engineering strategies for credit card fraud detection [J]. Expert Systems with Applications, 2016, 51: 134-142.

[3] HUANG S Y, LIN C C, CHIU A A, et al. Fraud detection using fraud triangle risk factors [J]. Information Systems Frontiers, 2017, 19(6): 1343-1356.

[4] SHUAIB M, OSHO O, ISMAILA I, et al. Comparative analysis of classification algorithms for email spam detection [J]. International Journal of Computer Network and Information Security, 2018, 10(1): 60.

[5] COLUCCIA A, DÁLCONZO A, RICCIATO F. Distribution-based anomaly detection via generalized likelihood ratio test: A general maximum entropy approach [J]. Computer Networks, 2013, 57(17): 3446-3462.

[6] VU N H, GOPALKRISHNAN V, ASSENT I. An Unbiased Distance-Based Outlier Detection Approach for High-Dimensional Data[C]// Database Systems for Advanced Applications - 16th International Conference (DASFAA 2011). Hong Kong, China, 2011.

[7] YU H, WANG B, XIAO G, et al. Distance-based outlier detection on uncertain data [J]. Journal of Computer Research & Development, 2010, 1(3): 293-298.

[8] RADOVANOVIC M, NANOPOULOS A, IVANOVIC M. Reverse nearest neighbors in unsupervised distance-based outlier detection [J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(5): 1369-1382.

[9] ZHANG Z, ZHU M, QIU J, et al. Outlier detection based on cluster outlier factor and mutual density[J]. International Journal of Intelligent Information and Database Systems, 2019, 12(1/2): 91-108.

[10] TANG B, HE H. A local density-based approach for outlier detection [J]. Neurocomputing, 2017, 241: 171-180.

[11] MISHRA G, AGARWAL S, JAIN P K, et al. Outlier detection using subset formation of clustering-based method[C]// International Conference on Advanced Computing Networking and Informatics. Singapore: Springer, 2019: 521-528.

[12] AZHAR F. Fuzzy clustering-based semi-supervised approach for outlier detection in big text data [J]. Progress in Artificial Intelligence, 2019, 8(1): 123-132.

[13] LI X, CHEN S. A Concise yet Effective model for Non-Aligned Incomplete Multi-view and Missing Multi-Label Learning [J]. arXiv: 2005. 00976, 2020.

[14] HU M, CHEN S. Doubly aligned incomplete multi-view clustering [J]. arXiv: 1903. 02785, 2019.

[15] WANG Z, XU J, CHEN S, et al. Regularized multi-view learning machine based on response surface technique [J]. Neurocomputing, 2012, 97: 201-213.

[16] QIAN Q, CHEN S, ZHOU X. Multi-view classification with cross-view must-link and cannot-link side information [J]. Knowledge-Based Systems, 2013, 54: 137-146.

[17] GAO J, FAN W, TURAGA D, et al. A spectral framework for detecting inconsistency across multi-source objects relationships [C]// 2011 IEEE 11-th International Conference on Data Mining. IEEE, 2011: 1050-1055.

[18] MARCOS A A, YAMADA M, KIMURA A, et al. Clustering-based anomaly detection in multi-view data[C]// Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 1545-1548.

[19] LI S, SHAO M, FU Y. Multi-view low-rank analysis for outlier detection[C]// Proceedings of the 2015 SIAM International Conference on Data Mining. SIAM, 2015: 748-756.

[20] ZHAO H, FU Y. Dual-regularized multi-view outlier detection [C]// Twenty-Fourth International Joint Conference on Artificial Intelligence. 2015.

[21] LI K, LI S, DING Z, et al. Latent discriminant subspace representations for multi-view outlier detection[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[22] LUO Y, TAO D, RAMAMOHANARAO K, et al. Tensor canonical correlation analysis for multi-view dimension reduction [J]. IEEE transactions on Knowledge and Data Engineering, 2015, 27(11): 3111-3124.

[23] CAO B, HE L, KONG X, et al. Tensor-based Multiview feature selection with applications to brain diseases[C]// 2014 IEEE International Conference on Data Mining. IEEE, 2014: 40-49.

[24] BU F. A high-order clustering algorithm based on dropout deep learning for heterogeneous data in cyber-physical-social systems [J]. IEEE Access, 2017, 6: 11687-11693.

[25] LI C G, VIDAL R. Structured sparse subspace clustering: A unified optimization framework[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 277-286.

[26] LIN Z, CHEN M, MA Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices [J]. arXiv: 1009. 5055, 2010.

[27] CAI J F, CANDÈS E J, SHEN Z. A singular value thresholding algorithm for matrix completion [J]. SIAM Journal on optimization, 2010, 20(4): 1956-1982.

[28] LIU G, LIN Z, YU Y. Robust subspace segmentation by low-rank representation[C]// Proceedings of the 27th International Conference on Machine Learning (ICML-10). DBLP, 2010: 663-670.

[29] LIU G, LIN Z, YAN S, et al. Robust recovery of subspace structures by low-rank representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 171-184.

ZHONG Ying-yu, born in 1995, postgraduate. His main research interests include multi-view learning and anomaly detection.

CHEN Song-can, born in 1962, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include pattern recognition, machine learning and neural computing.