Information-Aware Multi-View Outlier Detection

JINRONG LAI, School of Computer Science and Engineering, Sun Yat-sen University, China
TONG WANG, School of Computer Science and Engineering, Sun Yat-sen University, China
CHUAN CHEN*, School of Computer Science and Engineering, Sun Yat-sen University, China
ZIBIN ZHENG, School of Software Engineering, Sun Yat-sen University, China

With the development of multi-view learning, multi-view outlier detection has received increasing attention in recent years. However, the current research still faces two challenges: (1) The current research lacks theoretical analysis tools for multi-view outliers. (2) Most current multi-view outlier detection algorithms are based on shallow structural assumptions of the data, such as cluster assumptions and subspace assumptions, thus they are not suitable for more complex data distributions. In addressing these two issues, this paper proposes three occurrence mechanisms of multi-view outlier, which serve as foundational theoretical analysis tools for multi-view outliers. Utilizing proposed mechanisms, we analyze the impact of multi-view outliers and the information structure of multi-view data and validate our findings through experiments. Finally, we propose a novel algorithm referred to as Information-Aware Multi-View Outlier Detection (IAMOD). In contrast to other methods, IAMOD focuses on the information structure of multi-view data without relying on shallow structural assumptions. By learning a compact representation of the sample that is semantically rich and non-redundant, IAMOD can accurately identify multi-view outliers by comparing the consistency of the representations' neighbors and views. Extensive experimental results demonstrate that our approach outperforms several state-of-the-art multi-view outlier detection methods.

CCS Concepts: • **Computing methodologies** → **Anomaly detection**; *Neural networks.*

Additional Key Words and Phrases: Outlier Detection, Multi-view Learning, Information Theory

1 INTRODUCTION

Outlier detection is an important topic in data mining and machine learning. It is widely used in various fields, such as information security [9] and fault checking [3], fault detection [5]. However, these studies only focus on the data from one single view.

In many real-world scenarios, data often comes from multiple sources or heterogeneous data collected by multiple sensors. However, the detection of potential anomalous data in multi-view data has become a difficult problem. Currently, three types of outliers have been proposed, namely attribute outliers, class outliers, and class-attribute outliers, as shown in Figure 1. Class outliers are data samples that exhibit inconsistent feature behavior across different views. Attribute outliers are data samples that exhibit abnormal behaviors in some views. Class-attribute outlier exhibits class outlier characteristics in some views, while showing attribute outlier properties in the other views. After HOAD [6] first proposed the concept of horizontal outliers (class outliers),

*Corresponding author

Authors' addresses: Jinrong Lai, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China, 510006, laijr@mail2.sysu.edu.cn; Tong Wang, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China, 510006, wangt328@mail2.sysu.edu.cn; Chuan Chen, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, China, 510006, chenchuan@mail.sysu.edu.cn; Zibin Zheng, School of Software Engineering, Sun Yat-sen University, Zhuhai, Guangdong, China, 519082, zhzibin@mail.sysu.edu.cn.

ACM 1556-4681/2023/12-ART
https://doi.org/10.1145/3638354

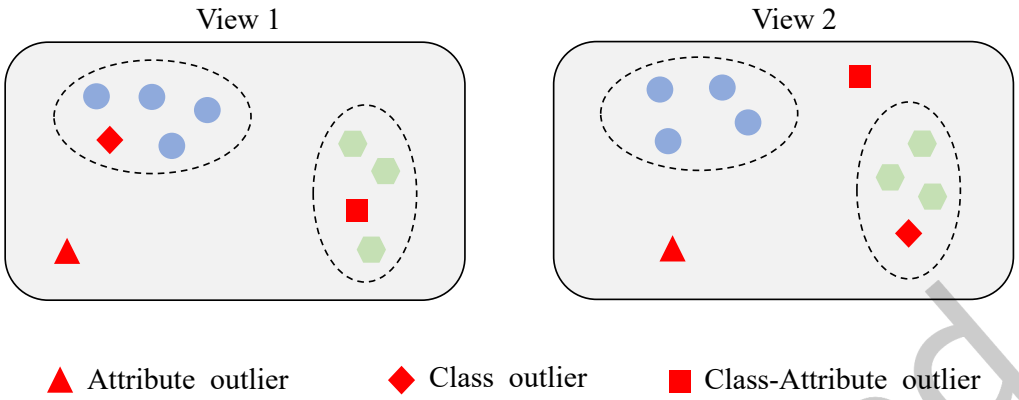

Fig. 1. Illustration of three types of multi-view outlier.

more and more work [4, 8, 10–12] began to focus on the detection of multi-view outliers. Among them, the two earlier proposed methods, AP [12] and HOAD, can only detect class outliers. Later, some methods [10, 11] based on subspace clustering appeared. These methods impose low-rank constraints on the data representation to learn a more robust multi-view representation that can detect multiple multi-view outliers simultaneously. However, a common disadvantage of these subspace clustering-based methods is the inability to quickly reason about new samples that have not been seen before. To address this problem, several neural network-based methods have been proposed recently. MODDIS [8] uses neural networks to integrate multi-view data into a potentially complete space in which outlier detection metrics are defined. NCMOD [4] is a state-of-the-art neural network-based method that learns low-dimensional encoding of raw data through a multi-view autoencoder, and then exploits the inter-view consistency of the low-dimensional encoding to detect multi-view outliers.

However, these current works only focus on multi-view outlier detection based on the behavioral characteristics of multi-view outliers and lack a theoretical analysis of multi-view outliers. Moreover, most of these current methods are based on the low-dimensional structure assumptions of data distribution, such as cluster structure, subspace structure, etc., thus they are not suitable for complex data distributions.

In this paper, we revisit outliers in these multi-view data and abstract three outlier occurrence mechanisms. Then, based on the above outlier occurrence mechanism, we study the impact of multi-view outliers on the information structure of multi-view data. Finally, based on multi-view information theory, we propose a novel multi-view outlier detection algorithm that can be applied to complex data distribution. Our major contributions are outlined as:

- We summarize three important multi-view outlier occurrence mechanisms. These occurrence mechanisms can summarize the most common multi-view outlier patterns and provide basic analysis tools for multi-view outlier research.
- Based on the proposed mechanisms of multi-view outlier occurrence, we theoretically demonstrate that multi-view outliers damage the semantic information shared between views and this is verified through experiments. This also indicates that the outlier occurrence mechanisms presented in this paper serve as a powerful tool for analyzing multi-view outliers.
- We propose a novel multi-view outlier detection algorithm based on the information theory,i.e., the Information-Aware Multi-View Outlier Detection (IAMOD). The algorithm aims to learn informative and compact semantic representations from raw data and detect the multi-view outliers based on these

semantic representations. Focusing on the information structure of multi-view data that is neglected by other methods, IAMOD is not reliant on shallow assumptions about the data distribution (such as cluster assumption and subspace clustering assumption) and can be applied to more complex data distributions.
- Extensive experiments demonstrate the effectiveness of IAMOD against state-of-the-art models. Even in the face of complex data distribution, IAMOD can still accurately detect multi-view outliers.

2 RELATED WORKS

In this section, we introduce two topics most relevant to our approach, including multi-view outlier detection and multi-view learning based on information theory.

2.1 Multi-View Outlier Detection

HOAD [6] is the pioneering work in multi-view outlier detection which proposed class outliers. Similarly, clustering-based algorithms such as AP [12] can also detect class outliers, while DMOD [17] can detect both attribute outliers and class outliers using k-means clustering. Latterly, low-rank analysis (MLRA [11] and LDSR [10]) have been proposed, with LDSR showing better performance on datasets with many views by learning a low-rank representation shared by all views. However, these methods can only handle datasets following the subspace assumption. To solve this problem, SRLSP [16] performs self-expression reconstruction by using only neighbors. The above methods based on spectral clustering or subspace clustering all face the out-of-sample problem: cannot perform inductive reasoning on new samples. Neural network-based methods do not have this flaw. MODDIS [8] uses neural networks to integrate multi-view data into a potentially complete space in which anomaly detection metrics are defined. NCMOD [4] is the latest neural network-based method and the best one at present. NCMOD exploits the neighbor consistency between representations from different views to detect multi-view outliers.

The common problem of all the above methods is that they ignore the exploration of the underlying mechanism of multi-view outliers and lack a theoretical analysis of multi-view outliers. Most of these methods rely on strict data distribution assumptions, so they cannot handle really complex data distributions.

2.2 Multi-view Learning Based on Information Theory

The earliest multi-view research based on information theory [14] successfully introduced information theory into the field of multi-view learning. DIM [7] first proposed the infomax principle, which maximizes the mutual information of the global representation and local representation of the sample. However, DIM does not focus on multi-view learning. Subsequently, inspired by DIM, the principle of maximizing mutual information is applied to multi-view learning problems and achieved success [1]. However, Infomin [15] believes that the representation should not contain too much redundant information, and proposes the Infomin principle to discard redundant information. Different from these works, we focus on studying the impact of multi-view outliers on the information structure of multi-view data and how to use the semantic information contained in the data to detect outliers.

3 ALGORITHM

In this section, we first establish the theoretical analysis basis of multi-view outliers. Then, we study the impact of multi-view outliers on the information structure of data according to these mechanisms. Finally, a novel multi-view outlier detection method based on information theory is proposed.

3.1 Theoretical Analysis Basis of Multi-view Outliers

3.1.1 Information Structure of Multi-view Data. In general, multi-view data is an attempt to describe and characterize an entity from different perspectives. No matter how different the data form of each view of a sample is (such as image view and text view), they are all descriptions of the same entity and will always share information about the entity. For example, for a dog, view 1 is the photo of the dog, and view 2 is the text describing the dog, that is, (*image of the dog*, *text about the dog*). The two views share some information (information about the dog), and these views have their own unique information (E.g, information about the background).

We note that the multi-view outliers will destroy the above-mentioned multi-view information structure. For example, for a multi-view outlier such as (*image of a dog*, *text about a cat*), its two views do not contain shared semantic information. Therefore, we can intuitively infer that the multi-view outliers will damage the mutual information shared between views. However, there is currently a lack of theoretical analysis research on multi-view outliers. To theoretically prove the above intuition, we need to explore the occurrence mechanism of multi-view outliers to help establish a theoretical basis for multi-view outliers.

3.1.2 Multi-view Outlier Occurrence Mechanisms. Due to the complex and diverse occurrence process of multi-view outliers in the real world, it is very difficult to conduct a relatively complete theoretical analysis. To address this challenge, we summarize three important multi-view outlier occurrence mechanisms that can cover most outlier generation processes:

Definition 3.1. **Random Mismatch**: The feature of view v of any sample is randomly replaced by the feature sampled from the data distribution of the view v according to a certain probability, which is called the random mismatch abnormality of view v.

The **random mismatch** mechanism covers outlier patterns due to the missing associativity of features across different views. For example, when some data from different views of multi-view data are independently collected and managed and lack alignment operations on samples, the resulting data association will be out of order. The random mismatch mechanism will produce class outliers.

Definition 3.2. **Targeted Tampering**: The data of view v is artificially tampered with so that the tampered data obeys the target data distribution, which is called targeted tampering of view v.

The **targeted tampering** mechanism models the outlier patterns caused by human tampering in the real world. For example, in a federated learning setting, if data from different views are stored in different client nodes, if one of these nodes is malicious, the malicious node's data tampering behavior is a kind of targeted tampering. The targeted tampering may produce class outliers, attribute outliers, and class-attribute outliers.

Definition 3.3. **Random Failure**: The samples in view v are replaced with random values or significant outliers with a certain probability, which is called a random failure of view v.

The **random failure** models the unavoidable outliers caused by natural causes in the process of data acquisition and storage in the real world. For example, significant outliers is introduced when a sensor that collects data fails or a failure occurs during data transmission. Random failure usually introduces attribute outliers that are far from the normal samples in each view.

The three multi-view outlier occurrence mechanisms we proposed cover the causes of most multi-view outliers and can be used as a basic analysis tool for multi-view outliers. In the next section, we will analyze the impact of multi-view outliers on shared semantic information across views based on these mechanisms.

3.2 Multi-view Outliers Decrease Semantic Information Shared across Views

By analyzing various multi-view outlier occurrence mechanisms, we will intuitively conclude that multi-view outliers will reduce semantic information shared across views. Next, we theoretically demonstrate this conclusion based on our proposed multi-view outlier occurrence mechanisms.

3.2.1 Theoretical Analysis. Here we theoretically prove the above-mentioned conclusion. We first consider the data distribution under ideal conditions without outliers. For brevity, we take multi-view data from two views as an example. It is assumed that there are c semantic classes in the multi-view data. Assuming without outliers occur, we can model the multi-view joint distribution as :

$$P_{\mathbb{V}_1\mathbb{V}_2}(i,j) = \begin{cases} p_i, 0 < p_i < 1 & (i = j \text{ and } i = 1, 2, ..., c) \\ 0, & (\text{other cases}) \end{cases} \tag{1}$$

where $P_{\mathbb{V}_1\mathbb{V}_2}(i,j)$ represents the probability of a sample that belongs to class i in view 1 and belongs to class j in view 2. For clarity, the joint distribution is recorded as Table 1. $P_{\mathbb{V}_1\mathbb{V}_2}(outlier, j)$ or $P_{\mathbb{V}_1\mathbb{V}_2}(i, outlier)$ represents the probability of a sample that appears as a significant outlier in a certain view. Normal samples correspond to the diagonal elements (excluding $P_{\mathbb{V}_1\mathbb{V}_2}(i = outlier, j = outlier)$) of Table 1.

Table 1. Ideal multi-view joint distribution without considering the occurrence of multi-view outliers

v_1 \ v_2	1	2	3	...	c	$outlier$
1	p_1	0	0	...	0	0
2	0	p_2	0	...	0	0
3	0	0	p_3	...	0	0
...	...	...	...	...	...	...
c	0	0	0	...	p_c	0
$outlier$	0	0	0	...	0	0

The shared semantic information in an ideal data distribution without outliers MI^{ideal} is:

$$\begin{aligned} MI^{ideal} &= \sum_{i,j} P_{\mathbb{V}_1\mathbb{V}_2}(i,j) log(\frac{P_{\mathbb{V}_1\mathbb{V}_2}(i,j)}{P_{\mathbb{V}_1}(i)P_{\mathbb{V}_2}(j)}) \\ &= \sum_{i\neq j} P_{\mathbb{V}_1\mathbb{V}_2}(i,j) log(\frac{P_{\mathbb{V}_1\mathbb{V}_2}(i,j)}{P_{\mathbb{V}_1}(i)P_{\mathbb{V}_2}(j)}) + \sum_{i=j} P_{\mathbb{V}_1\mathbb{V}_2}(i,j) log(\frac{P_{\mathbb{V}_1\mathbb{V}_2}(i,j)}{P_{\mathbb{V}_1}(i)P_{\mathbb{V}_2}(j)}) \\ &= 0 + \sum_{i=1}^{c} p_i log(\frac{p_i}{p_i \cdot p_i}) \\ &= -\sum_{i=1}^{c} p_i log(p_i) \end{aligned} \tag{2}$$

Next, based on our proposed multi-view outlier occurrence mechanisms, we demonstrate that multi-view outliers will lead to the reduction of semantic information shared across views.

(1) For random mismatch: Without loss of generality, we assume that random mismatch occurs in any one of the two views of the sample. Assuming that the probability of random mismatch occurrence is $0 < \alpha < 1$, then

the probability of random mismatch occurring in view 1 or view 2 are both $\alpha/2$. It is not difficult to calculate, the multi-view joint distribution considering random mismatch is:

$$\begin{aligned} P^{rm}_{\mathbb{V}_1\mathbb{V}_2}(i,j) =&(1-\alpha)P_{\mathbb{V}_1\mathbb{V}_2}(i,j) + \frac{\alpha}{2}P_{\mathbb{V}_2}(j)P_{\mathbb{V}_1\mathbb{V}_2}(i,i) \\ &+ \frac{\alpha}{2}P_{\mathbb{V}_1}(i)P_{\mathbb{V}_1\mathbb{V}_2}(j,j) \end{aligned} \tag{3}$$

Consider the multi-view marginal distribution under the random mismatch mechanism:

$$\begin{aligned} P^{rm}_{\mathbb{V}_1}(i) &= \sum_{j=1}^{c} P^{rm}_{\mathbb{V}_1\mathbb{V}_2}(i,j) = P_{\mathbb{V}_1}(i) \\ P^{rm}_{\mathbb{V}_2}(j) &= \sum_{i=1}^{c} P^{rm}_{\mathbb{V}_1\mathbb{V}_2}(i,j) = P_{\mathbb{V}_2}(j) \end{aligned} \tag{4}$$

The random mismatch mechanism does not change the marginal distribution of each view. This is also intuitive since random mismatch exceptions just destroy the association between views. Considering random mismatch, the mutual information between views MI^{rm} is:

$$\begin{aligned} MI^{rm} &= \sum_{i,j} P^{rm}_{\mathbb{V}_1\mathbb{V}_2}(i,j) log(\frac{P^{rm}_{\mathbb{V}_1\mathbb{V}_2}(i,j)}{P^{rm}_{\mathbb{V}_1}(i)P^{rm}_{\mathbb{V}_2}(j)}) \\ &= \sum_{i}[(1-\alpha)p_i + \alpha p_i^2] log(\frac{(1-\alpha)p_i + \alpha p_i^2}{p_i^2}) \\ &\quad + \alpha log\alpha \sum_{i\neq j} p_i p_j \\ &< \sum_{i}[(1-\alpha)p_i + \alpha p_i^2] log(\frac{(1-\alpha)p_i + \alpha p_i^2}{p_i^2}) \\ &< \sum_{i}[(1-\alpha)p_i + \alpha p_i] log(\frac{(1-\alpha)p_i + \alpha p_i}{p_i^2}) \\ &= -\sum_{i=1}^{c} p_i log(p_i) \\ &= MI^{ideal} \end{aligned} \tag{5}$$

So far, we have demonstrated that the outliers produced by random mismatch will lead to the reduction of semantic information shared across views.

(2) For targeted tampering: Artificial directed tampering attacks are complex and diverse. To simplify the problem, we consider a common abnormal situation: the attacker modifies a certain class of sample features in a view to another semantic class feature. Without loss of generality, we assume that the attacker tampered with the view 1 feature of samples belonging to semantic class 1 so that the view 1 feature belong to semantic class 2. Considering targeted tampering, the multi-view joint distribution is:

$$P^{tt}_{\mathbb{V}_1\mathbb{V}_2}(i,j) = \begin{cases} P_{\mathbb{V}_1\mathbb{V}_2}(i,j), & (i = j \text{ and } i = 2,...,c) \\ P_{\mathbb{V}_1\mathbb{V}_2}(1,1), & (i = 2 \text{ and } j{=}1) \\ 0, & (\text{other cases}) \end{cases} \tag{6}$$

Thus, the mutual information between views is:

$$\begin{aligned} MI^{tt} &= \sum_{i,j} P^{tt}_{\mathbb{V}_1\mathbb{V}_2}(i,j) log(\frac{P^{tt}_{\mathbb{V}_1\mathbb{V}_2}(i,j)}{P^{tt}_{\mathbb{V}_1}(i)P^{tt}_{\mathbb{V}_2}(j)}) \\ &= p_2 log(\frac{p_2}{(p_1+p_2)p_2}) + \sum_{i=3}^{k} p_i log(\frac{p_i}{p_i p_i}) \\ &\quad + p_1 log(\frac{p_1}{(p_1+p_2)p_1}) \\ &= p_1 log(\frac{p_1}{p_1+p_2}) + p_2 log(\frac{p_2}{p_1+p_2}) + MI^{ideal} \\ &< MI^{ideal} \end{aligned} \tag{7}$$

Therefore, we successfully demonstrate that the outliers produced by targeted tampering will lead to the reduction of semantic information shared across views.

(3) For random failure: We assume that the random failure rate of view 1 is α_1, and the random failure rate of view 2 is α_2. Considering the random failure, the multi-view joint distribution is:

$$P^{rf}_{\mathbb{V}_1\mathbb{V}_2}(i,j) = \begin{cases} (1-\alpha_1)(1-\alpha_2)P_{\mathbb{V}_1\mathbb{V}_2}(i,j), & (i,j\text{=1,2...c}) \\ \alpha_1(1-\alpha_2)P_{\mathbb{V}_2}(j), & (i\text{=outlier and } j\text{=1,2...c}) \\ \alpha_2(1-\alpha_1)P_{\mathbb{V}_1}(i), & (j\text{=outlier and } i\text{=1,2...c}) \\ \alpha_1\alpha_2, & (i\text{=outlier and } j\text{=outlier}) \end{cases} \tag{8}$$

We compute the semantic information shared across views according to the multi-view joint distribution :

$$\begin{aligned} MI^{rf} &= \sum_{i,j} P^{rf}_{\mathbb{V}_1\mathbb{V}_2}(i,j) log(\frac{P^{rf}_{\mathbb{V}_1\mathbb{V}_2}(i,j)}{P^{rf}_{\mathbb{V}_1}(i)P^{rf}_{\mathbb{V}_2}(j)}) \\ &= \sum_{i=1}^{c} (1-\alpha_1)(1-\alpha_2)p_i log[\frac{(1-\alpha_1)(1-\alpha_2)p_i}{(1-\alpha_1)p_i(1-\alpha_2)p_i}] \\ &\quad + \sum_{j=1}^{c} (1-\alpha_2)\alpha_1 p_j log(\frac{(1-\alpha_2)\alpha_1 p_j}{\alpha_1(1-\alpha_2)p_j}) \\ &\quad + \sum_{i=1}^{c} (1-\alpha_1)\alpha_2 p_i log(\frac{(1-\alpha_1)\alpha_2 p_i}{\alpha_2(1-\alpha_1)p_i}) \\ &\quad + \alpha_1\alpha_2 log(\frac{\alpha_1\alpha_2}{\alpha_1\alpha_2}) \\ &= (1-\alpha_1)(1-\alpha_2)MI^{ideal} \\ &\leqslant MI^{ideal} \end{aligned} \tag{9}$$

Thus, the multi-view outliers produced by random failure will lead to less semantic information shared between views. In summary, we demonstrate that the multi-view outlier will lead to a reduction of semantic information shared across views.

3.2.2 Experimental verification. We construct a series of outlier datasets with different types and proportions of outliers on the Caltech-7 dataset, and then estimate the sum of the mutual information between all view pairs according to the mutual information estimation method proposed by MINE [2] to get figure 2. As shown in the

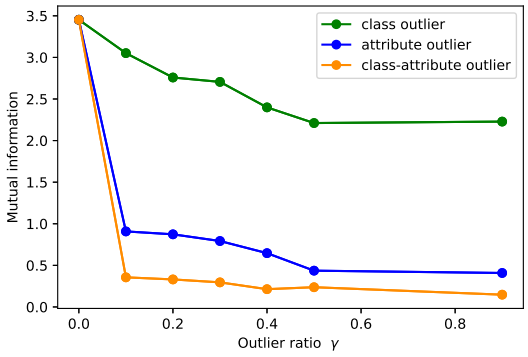

Fig. 2. Sum of mutual information values for all view pairs of the Caltech-7 dataset at different outlier ratios γ.

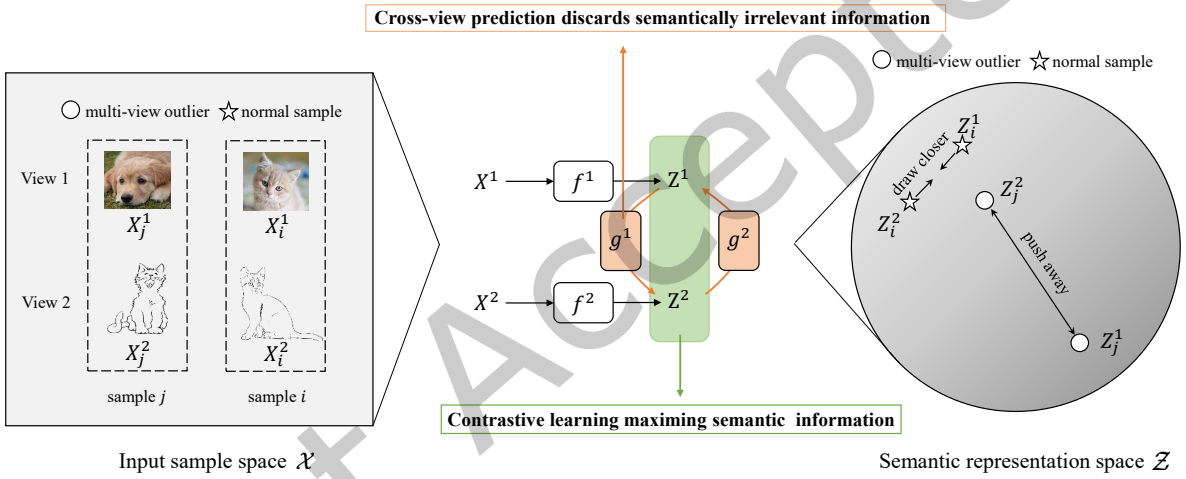

Fig. 3. Overview of the proposed IAMOD. We use contrastive learning to maximize the mutual information of representations from different views. Meanwhile, we use cross-view prediction to minimize the conditional entropy of representations from different views to discard semantically irrelevant information. Finally, we learn semantically rich and compact semantic representations for all samples.

figure 2, as the outlier ratio increases, the mutual information between views will decrease. This validates our previous analysis that outliers disrupt shared semantic information between views.

3.3 The Proposed Method

Inspired by the connection between multi-view outliers and the information structure of multi-view data, we propose IAMOD (as shown in Figure 3). IAMOD aims to learn representations containing precise semantic information so that the multi-view outliers with abnormal semantic information can be detected based on the learned semantic representations. The key to learning accurate semantic representations is to ensure that semantic representations contain as much semantic information as possible without semantically irrelevant information. To

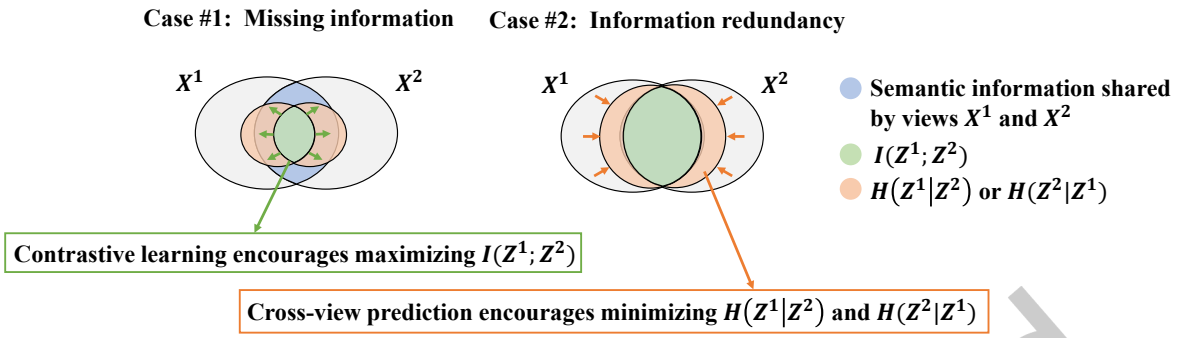

Fig. 4. Design motivation for the contrastive learning task and the cross-view prediction task

achieve this goal, we need to address two problems: (1) How to make the representation capture as much semantic information as possible? (2) How to make the representation discard semantically irrelevant information? To address these two problems, we separately design two tasks: the contrastive learning task and the cross-view prediction task (as shown in Figure 4).

3.3.1 Maximize semantic information shared by views. Take the case of two views as an example. When the learned representations Z^i lack semantic information, the contrastive learning task drives the representations Z^i to capture more semantic information, that is, the mutual information $I(Z^1; Z^2)$ between view 1 and view 2 (as shown in Figure 4). Specifically, we maximize the shared information of the representation Z^1 of view 1 and the representation Z^2 of view 2 by optimizing the contrastive loss:

$$\mathcal{L}_{CL} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{sim\left(\mathbf{Z}_i^1,\mathbf{Z}_i^2\right)}}{\sum_{j=1}^{N}e^{sim\left(\mathbf{Z}_i^1,\mathbf{Z}_j^2\right)}} \tag{10}$$

where N is the number of samples, $\mathrm{sim}(\cdot,\cdot)$ is euclidean distance, $\mathbf{Z}_\mathbf{i}^1$ is the representation of sample i corresponding to view 1, and so on. According to previous research [13], contrastive loss is related to mutual information : $\mathcal{L}_{CL} \geq \log(N) - I(Z^1; Z^2)$. Therefore, by minimizing the contrastive loss, we can force Z^1 and Z^2 to capture as much shared semantic information as possible.

3.3.2 Discard semantically irrelevant information. When the learned representation Z^i contains redundant semantically irrelevant information, the cross-view prediction task will force the representation Z^i to discard semantically irrelevant information $H(Z^1|Z^2)$ and $H(Z^2|Z^1)$ (as shown in the Figure 4).

Take minimizing conditional entropy $H(Z^1|Z^2)$ as an example. Minimizing $H(Z^1|Z^2)$ is equivalent to maximizing $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1 \mid Z^2\right)\right] = -H\left(Z^1 \mid Z^2\right)$. To avoid the intractable $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1 \mid Z^2\right)\right]$, we introduce a variational distribution $Q_\phi\left(Z^1 \mid Z^2\right)$ with ϕ as parameter. Notice:

$$\begin{aligned}
&\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1 \mid Z^2\right)\right] \\
&\quad = \max_{Q_\phi}\mathbb{E}_{P_{Z^1,Z^2}}\left[\log Q_\phi\left(Z^1 \mid Z^2\right)\right] \\
&\qquad + D_{\mathrm{KL}}\left(P\left(Z^1 \mid Z^2\right) \| Q_\phi\left(Z^1 \mid Z^2\right)\right) \\
&\quad \geq \max_{Q_\phi}\mathbb{E}_{P_{Z^1,Z^2}}\left[\log Q_\phi\left(Z^1 \mid Z^2\right)\right]
\end{aligned} \tag{11}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ means Kullback–Leibler divergence. Therefore, $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log Q_\phi\left(Z^1 \mid Z^2\right)\right]$ is a lower bound of $\mathbb{E}_{P_{Z^1,Z^2}}\left[\log P\left(Z^1 \mid Z^2\right)\right]$. For simplicity, we let $Q_\phi\left(Z^1 \mid Z^2\right)$ be the Gaussian distribution $\mathcal{N}\left(Z^1 \mid g^2(Z^2), \sigma\mathbf{I}\right)$. Then minimizing $H(Z^1|Z^2)$ is equivalent to minimizing $\mathcal{L}_{CP}^{1|2}$:

$$\begin{aligned}\mathcal{L}_{CP}^{1|2} =& \mathbb{E}_{g^2(Z^2),Z^1\sim P_{g^2(Z^2),Z^1}}\left[\left\|Z^1 - g^2(Z^2)\right\|_2^2\right] \\ =& \frac{1}{N}\sum_{i=1}^{N}\left\|Z_i^1 - g^2(Z_i^2)\right\|_2^2\end{aligned} \tag{12}$$

where $g^2(\cdot)$ is the predictor that predicts view 1 from view 2. Similarly, the loss function to achieve minimizing $H(Z^2|Z^1)$ is:

$$\begin{aligned}\mathcal{L}_{CP}^{2|1} =& \mathbb{E}_{g^1(Z^1),Z^2\sim P_{g^1(Z^1),Z^2}}\left[\left\|Z^2 - g^1(Z^1)\right\|_2^2\right] \\ =& \frac{1}{N}\sum_{i=1}^{N}\left\|Z_i^2 - g^1(Z_i^1)\right\|_2^2\end{aligned} \tag{13}$$

Therefore, the Cross-view prediction loss function is:

$$\begin{aligned}\mathcal{L}_{CP} =& \mathcal{L}_{CP}^{1|2} + \mathcal{L}_{CP}^{2|1} \\ =& \frac{1}{N}\sum_{i=1}^{N}\left[\left\|Z_i^2 - g^1(Z_i^1)\right\|_2^2 + \left\|Z_i^1 - g^2(Z_i^2)\right\|_2^2\right]\end{aligned} \tag{14}$$

3.3.3 Objective Function. The overall objective function of IAMOD is:

$$\mathcal{L} = \mathcal{L}_{CL} + \lambda\mathcal{L}_{CP} \tag{15}$$

where λ is hyper-parameter. This loss ensures that the representation captures as much semantic information as possible while discarding semantically irrelevant information.

3.3.4 Outlier Score Measurement. With the learned latent representations, we propose an outlier score function:

$$S(i) = S_{NC}(i) + S_{VC}(i) \tag{16}$$

where

$$\begin{aligned}S_{NC}(i) = \sum_{Z_j\in knn(Z_i)}\left[\left\|Z_i - Z_j\right\|_2^2\right], \\ (Z_i = Z_i^1 \oplus Z_i^2 \ldots \oplus Z_i^V)\end{aligned} \tag{17}$$

$$\begin{aligned}S_{VC}(i) = \sum_{v_1\neq v_2}[\|Z_i^{v_2} - g^{v_1\mapsto v_2}(Z_i^{v_1})\|_2^2 \\ +\|Z_i^{v_1} - g^{v_2\mapsto v_1}(Z_i^{v_2})\|_2^2]\end{aligned} \tag{18}$$

$knn(\cdot)$ is the k nearest neighbors. And the $g^{v_1\mapsto v_2}(\cdot)$ means the predictor that predicts view v_2 from view v_1. $S_{NC}(i)$ is the neighbor consistency score (concatenate all the representations Z_i^v of all views of sample i to get Z_i). $S_{VC}(i)$ is the view consistency score (cross-view prediction loss for the sample i). The outlier score function can detect all three types of outliers simultaneously: (1) For an attribute outlier i: Since it is a significant outlier in all views, it will have a larger neighbor consistency score. (2) For a class outlier i: Since it has inconsistent semantic information in different views, it will have a larger view consistency score. (3) For a class-attribute outlier i: Since it contains the characteristics of both attribute outlier and class outlier, it will have a larger neighbor consistency score and view consistency score.

Table 2. Datasets and related description: the table data is the feature dimension of each view (with feature name in parenthesis)

View	Synthetic dataset	UCI datasets					Real multi-view datasets		
		iris	pima	zoo	ionosphere	letter	MSRC-v1	AWA-10	Caltech-7
1	2	2	4	8	17	8	24 (CM)	2688 (CQ)	48 (Gabor)
2	2	2	4	8	17	8	576 (HOG)	2000 (LSS)	40 (WM)
3	-	-	-	-	-	-	512 (GIST)	252 (PHOG)	254 (CENTRIST)
4	-	-	-	-	-	-	256 (LBP)	2000 (SIFT)	1984 (HOG)
5	-	-	-	-	-	-	254 (CENT)	2000 (RGSIFT)	512 (GIST)
6	-	-	-	-	-	-	-	2000 (SURF)	928 (LBP)
Number of instances	400	150	768	101	351	1300	210	800	1474
Number of categories	2	3	2	7	2	26	7	10	7

4 EXPERIMENTS

4.1 Experimental Setup

4.1.1 Datasets. In this section, we use a synthetic dataset, UCI datasets, and three real multi-view datasets that are commonly used in multi-view learning. Details about the dataset are summarized in Table 2. In order to demonstrate the superiority of IAMOD in processing data that does not meet the cluster assumption, we construct a synthetic dataset. The synthetic dataset consists of a two-view dataset with one cluster structure but two categories, which does not meet the cluster structure assumption. View 1 is 400 data points sampled through a two-dimensional Gaussian distribution $N(0, 0, 0.1, 0.1, 0)$, and all sample points are divided into two categories by the straight line y=x. We perform a linear transformation on the data of view 1 and add random perturbation to obtain the data of view 2. The proportion of attribute outliers, class outliers, and class attribute outliers is all 5%. The proportion of attribute outliers, class outliers, and class attribute outliers is also all 5% in the real dataset.

All original datasets are free of outliers. Like most multi-view outlier detection work, we follow the approach of previous work [6] to generate multi-view outliers. To construct a certain proportion of attribute outliers, we randomly select some samples from the dataset and replace their features with random values. To construct a certain proportion of class outliers, we randomly select some sample pairs from the dataset and then swap the features for half of the views. To construct class-attribute outliers, we randomly select some sample pairs from the dataset, first swap the features of half of the views, and then replace the features of the other half of the views with random values.

4.1.2 Baselines. We compare the proposed method with six state-of-the-art multi-view outlier detection methods. HOAD [6] and AP [12] are cluster-based methods that can effectively detect class outliers. MLRA [11] and LDSR [10] are methods based on subspace self-expression theory, which can effectively detect three kinds of outliers simultaneously. MODDIS [8] and NCMOD [4] are state-of-the-art neural network-based methods that show superior performance in most cases.

4.1.3 Implementation details. We conducted our experiments on RTX 3090. The number of nearest neighbors in the outlier score function is set to 5. The trade-off parameter λ in the loss function is set to 10 (Synthetic), 100 (Caltech-7), 1 (AWA-10), and 10 (MSRC-v1) on different datasets. All encoders $f(\cdot)$ and inter-view predictors $g(\cdot)$ are implemented with a three-layer MLP(Multilayer Perceptron). The specific network structure is shown in Table 3. We use the Adam optimizer to optimize all encoders and predictors. The learning rate is set to 0.001 in all experiments. We follow the previous work [6, 8, 10, 11] and use AUC (area under ROC curve) as the evaluation metric. The source code is released at Github[1].

[1]https://github.com/MaybeLL/IAMOD

	MSRC-v1	AWA-10	Caltech-7
Encoder	FC(input_dim→8) BatchNormLayer ReLU FC(8→8) Softmax	FC(input_dim→128) BatchNormLayer ReLU FC(128→64) Softmax	FC(input_dim→128) BatchNormLayer ReLU FC(128→64) Softmax
Predictor	FC(8→8) BatchNormLayer ReLU FC(8→8) BatchNormLayer ReLU	FC(64→128) BatchNormLayer ReLU FC(128→64) BatchNormLayer ReLU	FC(64→128) BatchNormLayer ReLU FC(128→64) BatchNormLayer ReLU

Table 3. Network architecture details. "FC($x \rightarrow y$)" indicates a Fully Connected Layer with x input feature dimensions and y output feature dimensions.

Table 4. The comparison results on the synthetic dataset. AUC values (mean ± standard deviation) are reported. The best and the second-best results are in **bold** and underline, respectively.

	AUC (mean ± std)
HOAD	0.435±0.023
AP	0.672±0.038
MLRA	0.779±0.121
LDSR	0.824±0.061
MODDIS	0.972±0.018
NCMOD	0.881±0.053
IAMOD	**0.984±0.062**

4.2 Comparisons with State of the Arts

4.2.1 Synthetic Datasets. In order to demonstrate the superiority of IAMOD in processing data that does not meet the cluster assumption, we use a synthetic dataset to evaluate all methods. The experimental results are recorded in Table 4. Since IAMOD is constructed based on the semantic information structure of the data rather than the clustering structure, IAMOD can handle datasets that do not meet the cluster assumptions well. Likewise, MODDIS and NCMOD are not based on cluster assumption, so these two methods also perform significantly better than other methods.

4.2.2 UCI Datasets. To evaluate the detection ability of all methods for various multi-view outliers, we conduct experiments on the UCI dataset with two views. The experimental results are recorded in Tables 5, 6, 7. Experiments show that HOAD and AP can better detect class outliers, but perform poorly in attribute outlier detection. Because both methods are only designed to detect class outliers. Due to the small feature dimension and data size of these UCI datasets, neural network-based methods cannot fully exploit their full capabilities. Therefore, the neural network-based methods (MODDIS, NCMOD, and IAMOD) only show insignificant advantages compared with MLRA and LDSR. However, even so, IAMOD still outperforms other methods in most cases and can effectively detect various outliers.

Table 5. Evaluate the ability of various methods to detect attribute outliers. Comparison results on the UCI dataset with 10% attribute outliers. AUC values (mean ± standard deviation) are reported. The 1st/2nd best results are indicated in **bold**/underline.

	iris	pima	zoo	ionosphere	letter
HOAD	0.125±0.281	0.919±0.177	0.764±0.215	0.565±0.115	0.451±0.079
AP	0.039±0.046	0.024±0.016	0.416±0.165	0.480±0.092	0.445±0.028
MLRA	0.974±0.036	0.930±0.038	0.846±0.091	0.641±0.171	0.718±0.058
LDSR	0.976±0.049	0.987±0.011	0.932±0.042	0.729±0.067	0.997±0.001
MODDIS	0.740±0.195	0.714±0.027	0.896±0.018	0.703±0.033	0.834±0.061
NCMOD	0.983±0.012	**0.999±0.001**	0.867±0.051	0.619±0.031	0.998±0.001
IAMOD	**0.996±0.002**	0.997±0.002	**0.934±0.038**	**0.735±0.036**	**0.999±0.003**

Table 6. Evaluate the ability of various methods to detect class outliers. Comparison results on the UCI dataset with 10% class outliers. AUC values (mean ± standard deviation) are reported. The 1st/2nd best results are indicated in **bold**/underline.

	iris	pima	zoo	ionosphere	letter
HOAD	0.655±0.120	0.529±0.058	0.586±0.025	0.444±0.027	0.559±0.029
AP	**0.962±0.038**	0.497±0.031	**0.940±0.036**	**0.947±0.027**	0.837±0.015
MLRA	0.836±0.043	0.691±0.025	0.639±0.032	0.814±0.021	0.637±0.025
LDSR	0.750±0.087	0.637±0.035	0.824±0.061	0.833±0.013	0.752±0.027
MODDIS	0.816±0.082	0.697±0.037	0.784±0.069	0.789±0.039	0.857±0.020
NCMOD	0.591±0.141	0.533±0.035	0.839±0.052	0.841±0.032	0.548±0.023
IAMOD	0.934±0.024	**0.724±0.028**	0.903±0.067	0.863±0.031	**0.874±0.028**

4.2.3 Real Multi-View Datasets. To evaluate the ability of each method to handle complex data distribution, we conduct experiments on a real multi-view dataset with multiple views, and high dimensions. MLRA can only process those datasets with the same feature dimension per view, so it cannot participate in the comparison. The experimental results are recorded in Table 8. HOAD and AP perform very poorly compared to other methods. This is because both HOAD and AP depend on the constraints between pairs of views, and it is difficult to deal with a large number of views. Due to the small number of samples of MSRC-v1, it is difficult to train the neural network on this dataset. Therefore, on MSRC-v1, the subspace-based method LDSR performs slightly better than the neural network-based methods MODDIS, NCMOD and IAMOD. Since IAMOD is an algorithm based on the information structure of multi-view data, it does not depend on other strict data distribution assumptions. So on real complex multi-view datasets, IAMOD is superior to the method NCMOD based neighbor structure and the subspace-based method LDSR.

4.2.4 Ablation Study. In this part, we conduct the ablation study to further demonstrate the effectiveness of the proposed contrastive loss and cross-view prediction loss. The results of the ablation study are shown in Table 9. It

Table 7. Evaluate the ability of various methods to detect class-attribute outliers. Comparison results on the UCI dataset with 10% class-attribute outliers. AUC values (mean ± standard deviation) are reported. The 1st/2nd best results are indicated in **bold**/underline.

	iris	pima	zoo	ionosphere	letter
HOAD	0.448±0.155	0.354±0.084	0.740±0.028	0.429±0.067	0.248±0.111
AP	0.950±0.029	0.435±0.240	0.855±0.067	**0.902±0.010**	0.789±0.028
MLRA	0.879±0.032	0.746±0.022	0.815±0.054	0.722±0.034	0.654±0.031
LDSR	0.926±0.030	0.947±0.017	0.877±0.047	0.787±0.029	0.953±0.004
MODDIS	0.866±0.072	0.714±0.087	0.858±0.057	0.736±0.011	0.828±0.048
NCMOD	0.953±0.037	**0.975±0.010**	0.911±0.035	0.782±0.021	0.979±0.009
IAMOD	**0.962±0.033**	0.971±0.007	**0.917±0.016**	0.810±0.017	**0.999±0.001**

Table 8. The comparison results on the real multi-view datasets. AUC values (mean ± standard deviation) are reported. The best and the second-best results are in **bold** and underline, respectively.

	MSRC-v1	AWA-10	Caltech-7
HOAD	0.367±0.102	0.349±0.062	0.348±0.119
AP	0.512±0.035	0.461±0.026	0.459±0.041
LDSR	**0.971±0.011**	0.737±0.084	0.947±0.023
MODDIS	0.961±0.025	0.821±0.024	0.942±0.016
NCMOD	0.938±0.022	0.914±0.016	0.941±0.019
IAMOD	0.963±0.010	**0.939±0.013**	**0.954±0.008**

Table 9. The results of ablation experiments on the real multi-view datasets. AUC values (mean ± standard deviation) are reported. The best and the second-best results are in **bold** and underline, respectively.

	MSRC-v1	AWA-10	Caltech-7
L_{CL}	0.906±0.005	0.866±0.006	0.934±0.038
L_{CP}	0.886±0.042	0.799±0.005	0.881±0.012
$L_{CL} + L_{CP}$	**0.967±0.017**	**0.909±0.087**	**0.957±0.025**

can be seen that applying the contrastive learning task and the cross-view prediction task together outperforms applying one task alone, which demonstrates the effectiveness of our learning task design.

4.3 Parameter Analysis

Our method has two hyperparameters λ and k, where λ is the trade-off parameter in the loss function and k is the nearest neighbors number in the neighbor consistency score function. We conduct parameter analysis experiments on the Caltech-7 dataset and the results are recorded in Figure 5. We can observe that our method is fairly robust with various values of λ and k. Generally, we set λ to 100 and k to 6.

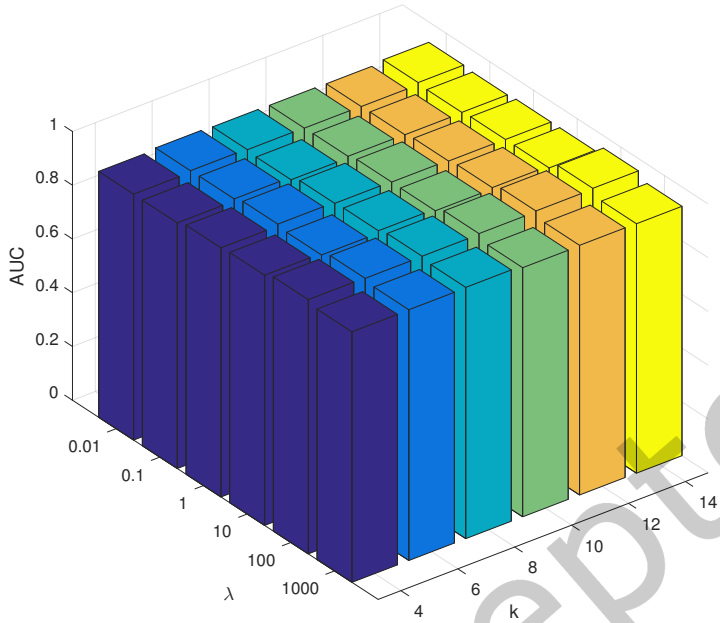

Fig. 5. Parameter analytical experiments on the Caltech-7 dataset.

5 CONCLUSION

We propose three significant multi-view outlier occurrence mechanisms that bridge the gap in theoretical multi-view outlier analysis. We then use these mechanisms to analyze the impact of multi-view outliers on the information structure of multi-view data. Based on this analysis, we introduce a novel multi-view outlier detection algorithm grounded in information theory. Our experiments demonstrate the superior performance of IAMOD, which can accurately detect even the most complex multi-view outliers. We hope that our work will stimulate more researchers to take an interest in multi-view outlier detection. Our approach provides a perspective based on information theory. In the future, we plan to focus on addressing the problem of multi-view outlier detection in supervised or semi-supervised scenarios. We will also explore the relationship between information bottleneck theory and this problem.

ACKNOWLEDGMENTS

The research is supported by the National Key Research and Development Program of China (2023YFB2703700), the Key-Area Research and Development Program of Shandong Province (2021CXGC010108), the National Natural Science Foundation of China (62176269), the Guangzhou Science and Technology Program (2023A04J0314)

REFERENCES

[1] Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems* 32 (2019).

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*. PMLR, 531–540.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–58.
[4] Li Cheng, Yijie Wang, and Xinwang Liu. 2021. Neighborhood Consensus Networks for Unsupervised Multi-view Outlier Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7099–7106.
[5] Qi Ding, Natallia Katenka, Paul Barford, Eric Kolaczyk, and Mark Crovella. 2012. Intrusion as (anti) social communication: characterization and detection. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 886–894.
[6] Jing Gao, Wei Fan, Deepak Turaga, Srinivasan Parthasarathy, and Jiawei Han. 2011. A spectral framework for detecting inconsistency across multi-source object relationships. In *2011 IEEE 11th International Conference on Data Mining*. IEEE, 1050–1055.
[7] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).
[8] Yu-Xuan Ji, Ling Huang, Heng-Ping He, Chang-Dong Wang, Guangqiang Xie, Wei Shi, and Kun-Yu Lin. 2019. Multi-view outlier detection in deep intact space. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1132–1137.
[9] Zhao Kang, Yiwei Lu, Yuanzhang Su, Changsheng Li, and Zenglin Xu. 2019. Similarity learning via kernel preserving embedding. In *proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4057–4064.
[10] Kai Li, Sheng Li, Zhengming Ding, Weidong Zhang, and Yun Fu. 2018. Latent discriminant subspace representations for multi-view outlier detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[11] Sheng Li, Ming Shao, and Yun Fu. 2018. Multi-view low-rank analysis with applications to outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 3 (2018), 1–22.
[12] Alejandro Marcos Alvarez, Makoto Yamada, Akisato Kimura, and Tomoharu Iwata. 2013. Clustering-based anomaly detection in multi-view data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 1545–1548.
[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
[14] Karthik Sridharan and Sham M Kakade. 2008. An information theoretic framework for multi-view learning. (2008).
[15] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems* 33 (2020), 6827–6839.
[16] Yu Wang, Chuan Chen, Jinrong Lai, Lele Fu, Yuren Zhou, and Zibin Zheng. 2022. A Self-Representation Method with Local Similarity Preserving for Fast Multi-View Outlier Detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* (2022).
[17] Handong Zhao, Hongfu Liu, Zhengming Ding, and Yun Fu. 2017. Consensus regularized multi-view outlier detection. *IEEE Transactions on Image Processing* 27, 1 (2017), 236–248.