Multi-source Manifold Outlier Detection

Lei Zhang[1](✉), Shupeng Wang[1](✉), Ge Fu[2](✉), Zhenyu Wang[1], Lei Cui[1],
and Junteng Hou[1]

[1] Institute of Information Engineering, CAS, Beijing 100093, China
{zhanglei1,wangshupeng,wangzhenyu,cuilei,houjunteng}@iie.ac.cn
[2] CNCERT/CC, Beijing 100029, China
fg@cert.org.cn

Abstract. Outlier detection is an important task in data mining, with many practical applications ranging from fraud detection to public health. However, with the emergence of more and more multi-source data in many real-world scenarios, the task of outlier detection becomes even more challenging as traditional mono-source outlier detection techniques can no longer be suitable for multi-source heterogeneous data. In this paper, a general framework based the consistent representations is proposed to identify multi-source heterogeneous outlier. According to the information compatibility among different sources, Manifold learning are combined in the proposed method to obtain a shared representation space, in which the information-correlated representations are close along manifold while the semantic-complementary instances are close in Euclidean distance. Furthermore, the multi-source outliers can be effectively identified in the affine subspace which is learned through affine combination of shared representations from different sources in the feature-homogeneous space. Comprehensive empirical investigations are presented that confirm the promise of our proposed framework.

Keywords: Multi-source · Manifold learning · Heterogeneous · Outlier detection

1 Introduction

Recent years have witnessed significant advances in multi-source learning. Many multi-source techniques have been developed to assist people in extracting useful information from rapidly growing volumes of multi-source data [25,26]. However, unlike the previous work that focused on the study of mining general patterns from different sources, it is worth to note that no existing efforts have focused on the detection of multi-source heterogeneous outliers. In particular, this detection is generally performed to identify abnormal heterogeneous observation from different sources.

Furthermore, detecting outliers is more interesting and useful than identifying normal instances [4,11,17]. For example, to protect the properties of customers, an electronic commerce detection system can monitor the customers' financial

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11537, pp. 44–58, 2019.
https://doi.org/10.1007/978-3-030-22741-8_4

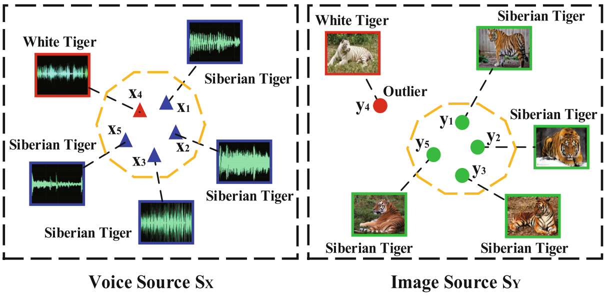

Fig. 1. Multi-source heterogeneous outlier

activities in order to identify abnormal consuming behavior of credit card as criminal activities (outlier). Consequently, multi-source outlier detection is an important task in data mining, with many practical applications ranging from fraud detection to public health. Many outlier detection methods have been proposed over the past decades, including mono-source [6,12,13,18] and multi-view [10,14,21,27] outlier detection.

Mono-source Outlier Detection. Recently, some researchers have investigated many machine learning methods [6,12,13,18] to deal with outlier detection problems in mono-source data. Knorr et al. pointed out in [12] that the identification of outliers can lead to the discovery of truly unexpected knowledge in various actual application fields. Meanwhile, they proposed and analyzed several algorithms for finding DB-outliers. In [13], Li et al. presented a representation-based method to calculate the reverse unreachability of a point to evaluate to what degree this observation is boundary point or outlier. Rahmani and Atia [18] proposed two randomized algorithms for two distinct outlier models namely, sparse and independent outlier models based robust principal component analysis. k-Nearest Neighbors (kNN) [6] defines the distance of a given data point to its kth nearest neighbors as the outlier. The greater the value of the score, the most likely the outlier.

However, with the emergence of more and more multi-source data in many real-world scenarios, the task of outlier detection becomes even more challenging as traditional mono-source outlier detection techniques can no longer be suitable for multi-source heterogeneous data. For example, as shown in Fig. 1, it is greatly difficult in the single voice source to discovery the outlier x_4, i.e., White tiger, hidden in the cluster of Siberian tiger. However, according to the distribution of the corresponding Image source, the heterogeneous outlier y_4 can be easily identified. Thus, it is necessary to develop an effective outlier detection method for multi-source heterogeneous data.

Multi-view Outlier Detection. To overcome the drawback of traditional mono-source methods, several efforts [10,14,21,27] have been devoted to identifying multi-source outliers. In [21], Li et al. proposed a multi-view outlier detection framework to detect two different types of outliers from multiple views simultaneously. The characteristic of this method is that it only can detect the outliers from different views in their own original low-level feature spaces. Similarly, Zhao and Fu [27] also investigated multi-view outlier detection problem to detect two different kinds of anomalies simultaneously. Different from Li's approach, they represented both kinds of outliers in two different spaces, i.e. latent space and original feature space. Janeja and Palanisamy presented in [10] a two-step method to find anomalous points across multiple domains. This technique first conducted single-domain anomaly detection to discover outliers in each domain, then mined association rule across domains to discover relationship between anomalous points. A multi-view anomaly detection algorithm was developed by Liu and Lam in [14] to find potentially malicious insider activity across multiple data sources.

Difficulties and Challenges. Generally, these existing methods tend to identify multi-source outliers from different feature spaces by using association analysis across sources. Most of these methods are, however, designed for detecting the Class-outliers [21,27] in spaces of the original attributes, and discovering Attribute-outliers [21,27] in combinations of underlying spaces. Consequently, these methods will face an enormous challenge in the real-world applications for the following reason. Due to the attempt of identifying multi-source outliers in different original feature spaces, it is extremely difficulty for the above-mentioned approaches to capture much more complementary information from different sources. It will lead to a low recognition rate for multi-source outliers. Furthermore, it has been proved in [26] that the consistent representations for multi-source heterogeneous data will be more favorable for fully exploiting the complementarity among different sources. Thus, it is inevitably an urgent problem to detect all kinds of multi-source outliers from different sources in a consistent feature-homogeneous space.

1.1 Main Contributions

The key contributions of this paper are highlighted as follows:

- To detect multi-source heterogeneous outliers, a general Multi-Source Manifold Outlier Detection (MMOD) framework based the consistent representations for multi-source heterogeneous data is proposed.
- Manifold learning is integrated in the framework to obtain a shared-representation space, in which the information-correlated representations are close along manifold while the semantic-complementary instances are close in Euclidean distance.
- According to the information compatibility among different sources, an affine subspace is learned through affine combination of shared representations from different sources in the feature-homogeneous space.
- Multi-source heterogeneous outliers can be effectively identified in the affine subspace under the constraints of information compatibility among different sources.

1.2 Organization

The remainder of this paper is organized as follows: In Sect. 1.3, the notations are formally defined. We present a general framework for detecting multi-source heterogeneous outliers in Sect. 2.1. Furthermore, Sect. 2.2 provides an efficient algorithm to solve the proposed framework. Experimental results and analyses are reported in Sect. 3. Section 4 concludes this paper.

1.3 Notations

In this section, some important notations are summarized into Table 1 for convenience.

Table 1. Notations

Notation	Description
S_x	Source X
S_y	Source Y
$X_N \in \mathbb{R}^{n_1 \times d_x}$	Normal samples in S_x
$Y_N \in \mathbb{R}^{n_1 \times d_y}$	Normal samples in S_y
$X_S \in \mathbb{R}^{n_2 \times d_x}$	Suspected outliers in S_x
$x_i \in \mathbb{R}^{d_x}$	The i-th sample from S_x
$y_i \in \mathbb{R}^{d_y}$	The i-th sample from S_y
n_1	Number of normal data
n_2	Number of suspected outliers
d_x	Dimensionality of S_x
d_y	Dimensionality of S_y
(x_i, y_i)	The i-th multi-source datum
$\|\| \cdot \|\|_F$	Frobenius norm
$\triangledown f(\cdot)$	Gradient of smooth function $f(\cdot)$

2 Detecting Multi-source Heterogeneous Outliers

Here we propose a general framework to detect heterogeneous outliers in multi-source datasets.

2.1 The Proposed MMOD Model

In the light of the existing multi-source methods' shortcomings, we focus on a particularly important problem of identifying all kinds of multi-source heterogeneous outliers in a consistent feature-homogeneous space.

In general, outliers are located around the margin of the data set with high density, such as a cluster. Furthermore, Elhamifar [7] has pointed out that each

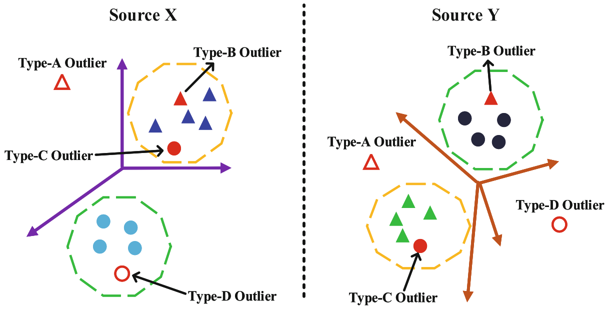

Fig. 2. 4 types of heterogeneous outliers.

data point in a union of subspaces can be efficiently represented as a linear or affine combination of other points in the dataset. Meanwhile, Bertsekas has also proved that all convex combinations are geometrically within the convex hull of the given points [2]. Consequently, the negative components in the representation correspond to the outliers outside the convex combination of its neighbors [5,20].

Following the above-mentioned theoretical results, we propose a novel Multi-Source Manifold Outlier Detection (MMOD) framework based the consistent representations for multi-source heterogeneous data. The main goal of the proposed framework is unified detection of outliers from heterogeneous datasets in a feature-homogeneous space, in order to avoid wasting the complementary information among different sources and improve the recognition rate for multi-source outliers.

In this paper, we focus on detecting all kinds of heterogeneous outliers (See Fig. 2) from multi-source heterogeneous data. In particular, we aim to identify four types of heterogeneous outliers that are defined below.

Definition 1. Type-A outliers have consistent abnormal behaviors in each source.

Definition 2. Type-B outliers are deviant instances that show normal clustering results in one source but abnormal cluster memberships in another source.

Definition 3. Type-C outliers own abnormal clustering results in each source.

Definition 4. Type-D outliers refer to exceptional samples that exhibit normal clustering results in one source but abnormal behavior in another source.

Given a normal dataset $X_N = \{x_1, x_2, \cdots, x_{n_1}\} \in \mathbb{R}^{d_x \times n_1}$ and a sample $c \in \mathbb{R}^{d_x}$, an affine space $H = \{w \in \mathbb{R}^{n_1} | X_N w = c\}$ can be spanned by its neighbors from Source X. Note that w is the representation of c in the affine space, and w_i is the component of the representation w of c. Generally, outliers are located around the margin of the dataset with high density, such as a cluster.

It is known that if $0 \leq w_i \leq 1$, then the point c will be within (or on the boundary) of the convex hull. If any w_i is less than zero or greater than 1, then the point will lie outside the convex hull. Thus, data representation can uncover the intrinsic data structure. Obviously, heterogeneous outliers can be identified according to the following principle.

1. c is normal point if $0 \leq w_i \leq 1$.
2. c is abnormal point (outlier) if any $w_i < 0$ or $w_i > 1$.

Specifically, the new distance metrics are defined as follows to learn a Mahalanobis distance [23]:

$$\mathcal{D}_{M_X}(x_i, x_j) = (x_i - x_j)^T M_X (x_i - x_j), \tag{1}$$

$$\mathcal{D}_{M_Y}(y_i, y_j) = (y_i - y_j)^T M_Y (y_i - y_j), \tag{2}$$

where $M_X = A^T A$ and $M_Y = B^T B$ are two positive semi-definite matrices. Thus, the linear transformations A and B can be applied to each pair of co-occurring heterogeneous representations (x_i, y_i).

Then the proposed approach can be formulated as follows:

$$\boldsymbol{\Psi_1} : \begin{array}{ll} \underset{A,B,M,W}{min} & \| X_N A M B^T Y_N^T \|_F^2 + \alpha \| X_N A - Y_N B \|_F^2 \\ s.t. & A^T A = I \quad and \quad B^T B = I \quad and \quad M \succeq 0 \\ & X_S A = W^T Y_N B \end{array} \tag{3}$$

where $A \in \mathbb{R}^{d_x \times k}$, $B \in \mathbb{R}^{d_y \times k}$, k is the dimensionality of the feature-homogeneous subspace, and α is a trade-off parameter. The first item of the objective function in the model Ψ_1 is to measure the smoothness between different linear transformations A and B to extract the information correlation among heterogeneous representations. Moreover, the motivation of introducing the second item in the objective function is to capture the semantic complementarity among different sources. The orthogonal constraints $A^T A = I$ and $B^T B = I$ are added into the optimization to effectively remove the correlations among different features in the same source, the positive semidefinite restraint $M \in \mathbb{S}_+^{k \times k} \succeq 0$ can ensure a well-defined pseudo-metric. To identify multi-source heterogeneous outliers, the affine hull constraint $X_S A = W^T Y_N B$ based on data representation is added into the model Ψ_1 to learn an affine subspace. The matrices $W \in \mathbb{R}^{n_1 \times n_2}$ encodes the neighborhood relationships between points in the affine subspace, $w_i \in \mathbb{R}^{n_1}$ is the representation of $x_S^i \in \mathbb{R}^{d_x}$ in the affine subspace, respectively.

Note that solving the problem Ψ_1 in Eq. (3) directly is a challenging task for two main reasons. First, it is difficult to seek the solution that satisfies the convex hull constraint. Second, the orthogonal constraints are not smooth, which makes it even more difficult to compute the optimum. Thus, we propose to use Lagrangian duality to augment the objective function with a weighted sum of the convex hull constraint to obtain a solvable problem Ψ_2 as follows:

$$\boldsymbol{\Psi_2} : \begin{array}{l} \underset{A,B,M,W}{min} \| X_N A M B^T Y_N^T \|_F^2 + \alpha \| X_N A - Y_N B \|_F^2 + \beta \| X_S A - W^T Y_N B \|_F^2 \\ s.t. \quad A^T A = I \quad and \quad B^T B = I \quad and \quad M \succeq 0 \end{array} \tag{4}$$

In Sect. 2.2, an efficient algorithm is proposed to solve the problem Ψ_2.

2.2 An Efficient Solver for Ψ_2

Here we provide an efficient algorithm to solve Ψ_2.

The optimization problem Ψ_2 in Eq. (4) can be simplified as follows:

$$\underset{Z \in \mathcal{C}}{min} \quad F(Z) = \| \cdot \|_F + \alpha \| \cdot \|_F + \beta \| \cdot \|_F, \tag{5}$$

where $F(\cdot)$ is a smooth objective function, $Z = [A_Z \quad B_Z \quad M_Z \quad W_Z]$ symbolically represents the optimization variables, and $\mathcal{C}$ is the closed domain with respect to each variable:

$$\mathcal{C} = \{Z | A_Z^T A_Z = I, B_Z^T B_Z = I, M_Z \succeq 0\}. \tag{6}$$

Obviously, the optimization problem in Eq. (5) is non-convex. However, Ando and Zhang have testified in [1] that the alternating optimization method can effectively solve non-convex problem. They have also pointed out that this method usually did not lead to serious problems since given the local optimal solution of one variable, the solution of other variables would still be globally optimal.

Additionally, the problem in Eq. (5) is separately convex with respect to each optimization variable. Furthermore, as $F(\cdot)$ is continuously differentiable with Lipschitz continuous gradient [15] with respect to each variable, respectively. Thus, through combining Accelerated Projected Gradient (APG) [15] method and alternating optimization approach [1], the problem in Eq. (5) can be effectively solved.

However, the non-convex optimization problem in Eq. (5) is generally difficult to optimize due to the orthogonal constraints. Guo and Xiao have pointed out in [8] that Gradient Descent Method with Curvilinear Search (GDMCS) in [24] can effectively solve non-convex optimization problem for a local optimal solution as long as the Armijo-Wolfe conditions are satisfied.

Furthermore, since the objective function in Eq. (5) is smooth, the gradient of the objective function with respect to A, B can be easily computed, respectively. In each iteration of the gradient descent procedure, given the current feasible point (A, B), the gradients can be computed as follows:

$$G_1 = \nabla_A F(A, B), \tag{7}$$

$$G_2 = \nabla_B F(A, B). \tag{8}$$

We then compute two skew-symmetric matrices:

$$F_1 = G_1 A^T - A G_1^T, \tag{9}$$

$$F_2 = G_2 B^T - B G_2^T. \tag{10}$$

It is easy to see $F_1^T = -F_1$ and $F_2^T = -F_2$. The next new point can be searched as a curvilinear function of a step size variable τ, such that

$$Q_1(\tau) = (I + \tau F_1/2)^{-1}(1 - \tau F_1/2)A, \tag{11}$$

$$Q_2(\tau) = (I + \tau F_2/2)^{-1}(1 - \tau F_2/2)B. \tag{12}$$

It is easy to verify that $Q_1(\tau)^T Q_1(\tau) = I$ and $Q_2(\tau)^T Q_2(\tau) = I$ for all $\tau \in \mathbb{R}$. Thus we can stay in the feasible region along the curve defined by τ. Moreover, $dQ_1(0)/d\tau$ and $dQ_2(0)/d\tau$ are equal to the projections of $(-G_1)$ and $(-G_2)$ onto the tangent space $\mathcal{C}$ at the current point (A, B). Hence $\{Q_1(\tau), Q_2(\tau)\}_{(\tau \geq 0)}$ is a descent path in the close neighborhood of the current point. We thus apply a similar strategy as the standard backtracking line search to find a proper step size τ using curvilinear search, while guaranteeing the iterations to converge to a stationary point. We determine a proper step size τ as one satisfying the following Armijo-Wolfe conditions [24]:

$$F(Q_1(\tau), Q_2(\tau)) \leq F(Q_1(0), Q_2(0)) + \rho_1 \tau F'_\tau(Q_1(0), Q_2(0)), \tag{13}$$

$$F'_\tau(Q_1(\tau), Q_2(\tau)) \geq \rho_2 F'_\tau(Q_1(0), Q_2(0)). \tag{14}$$

Here $F'_\tau(Q_1(\tau), Q_2(\tau))$ is the derivative of F with respect to τ,

$$\begin{aligned} &F'_\tau(Q_1(\tau), Q_2(\tau)) = \\ &-tr((\nabla_A F(Q_1(\tau), Q_2(\tau)))^T (I + \frac{\tau}{2} F_1)^{-1} F_1 (\frac{A + Q_1(\tau)}{2}) \\ &-tr((\nabla_B F(Q_1(\tau), Q_2(\tau)))^T (I + \frac{\tau}{2} F_2)^{-1} F_2 (\frac{B + Q_2(\tau)}{2}). \end{aligned} \tag{15}$$

Therefore,

$$\begin{aligned} F'_\tau(Q_1(0), Q_2(0)) &= -tr(G_1^T (G_1 A^T - A G_1^T) A) \\ &\quad - tr(G_2^T (G_2 B^T - B G_2^T) B) \\ &= -\frac{\|F_1\|_F^2}{2} - \frac{\|F_2\|_F^2}{2}. \end{aligned} \tag{16}$$

Accordingly, it is appropriate to use the gradient descent method to solve the problem Ψ_2 in Eq. (5).

The APG algorithm is a first-order gradient method, which can accelerate each gradient step on the feasible solution to obtain an optimal solution when minimizing a smooth function [16]. This method will construct a solution point sequence $\{Z_i\}$ and a searching point sequence $\{S_i\}$, where each Z_i is updated from S_i.

Furthermore, a given point s in the APG algorithm needs to be projected into the set $\mathcal{C}$:

$$proj_{\mathcal{C}}(s) = arg \min_{z \in \mathcal{C}} \|z - s\|_F^2 / 2. \tag{17}$$

Weinberger et al. proposed a Positive Semi-definite Projection (PSP) [23] to minimize a smooth function while remaining positive semi-definite constraints. It will project optimal variables into a cone of all positive semi-definite matrices after each gradient step. The projection is computed from the diagonalization of optimal variables, which effectively truncates any negative eigenvalues from the gradient step, setting them to zero. Then we can use the PSP to solve the problem in Eq. (17).

Finally, to solve the problem in Eq. (5), the projection $Z = [A_Z \ \ B_Z \ \ M_Z \ \ W_Z]$ of a given point $S = [A_S \ \ B_S \ \ M_S \ \ W_S]$ onto the set $\mathcal{C}$ is defined by:

$$proj_{\mathcal{C}}(S) = arg \min_{Z \in \mathcal{C}} \|Z - S\|_F^2 / 2. \tag{18}$$

By combining APG, GDMCS, and PSP, we can solve the problem in Eq. (18). The overall algorithm is given in Algorithm 1, where the function $\boldsymbol{Schmidt}(\cdot)$ denotes the GramSchmidt process.

Algorithm 1. Multi-Source Manifold Outlier Detection (**MMOD**)

```
Input: $F(\cdot)$, $Z_0 = [A_{Z_0} \ B_{Z_0} \ M_{Z_0} \ W_{Z_0}]$, $\tau_1$, $0 < \rho_1 < \rho_2 < 1$, $s=1$, $t=1$, $p=1$,
       $q=1$, $\eta_1 > 0$, $\theta_1 > 0$, $\lambda_1 > 0$, $\mu_1 > 0$.
Output: $Z^*$.
Set $A_{Z_1} = A_{Z_0}$, $B_{Z_1} = B_{Z_0}$, $M_{Z_1} = M_{Z_0}$, and $W_{Z_1} = W_{Z_0}$.
for $i$ =1,2,$\cdots$,$max\text{–}iter$ do
    Fix $B$, $M$, $W$ and approximately solve for $A$.
    Define $F_{\eta, A_S}(A_Z) = F(A_S) + \langle \nabla F(A_S), A_Z - A_S \rangle + \eta \|A_Z - A_S\|_F^2 / 2$.
    for $j$ =1,2,$\cdots$, $h_1$ do
        Set $a_j = (s-1)/s$.
        Compute $A_{S_i} = (1 + \alpha_j) A_{Z_i} - \alpha_j A_{Z_{i-1}}$.
        Compute $\nabla_{A_S} F(A_{S_i})$.
        while (true)
            Compute $\widehat{A_S} = A_{S_i} - \nabla_{A_S} F(A_{S_i}) / \eta_i$.
            Compute $[\widehat{A_S}] = \boldsymbol{Schmidt}(\widehat{A_S})$.
            Set $[A_{Z_{i+1}}] = \boldsymbol{GDMCS}(F(\cdot), \widehat{A_S}, \tau_1, \rho_1, \rho_2)$.
            if $F(A_{Z_{i+1}}) \leq F_{\eta_i, A_{S_i}}(A_{Z_{i+1}})$, then break;
            else Update $\eta_i = \eta_i \times 2$.
            end-if
         end-while
         Update $s = \left(1 + \sqrt{1 + 4s^2}\right)/2$, $\eta_{i+1} = \eta_i$.
    end-for
    Fix $A$, $M$, $W$ and approximately solve for $B$.
    Define $F_{\theta, B_S}(B_Z) = F(B_S) + \langle \nabla F(B_S), B_Z - B_S \rangle + \theta \|B_Z - B_S\|_F^2 / 2$.
    for $j$ =1,2,$\cdots$, $h_2$ do
        Set $a_j = (t-1)/t$.
        Compute $B_{S_i} = (1 + \alpha_j) B_{Z_i} - \alpha_j B_{Z_{i-1}}$.
        Compute $\nabla_{B_S} F(B_{S_i})$.
        while (true)
            Compute $\widehat{B_S} = B_{S_i} - \nabla_{B_S} F(B_{S_i}) / \theta_i$.
            Compute $[\widehat{B_S}] = \boldsymbol{Schmidt}(\widehat{B_S})$.
            Set $[B_{Z_{i+1}}] = \boldsymbol{GDMCS}(F(\cdot), \widehat{B_S}, \tau_1, \rho_1, \rho_2)$.
            if $F(B_{Z_{i+1}}) \leq F_{\theta_i, B_{S_i}}(B_{Z_{i+1}})$, then break;
            else Update $\theta_i = \theta_i \times 2$.
            end-if
```

```
      end-while
      Update $t = \left(1+\sqrt{1+4t^2}\right)/2$, $\theta_{i+1}=\theta_i$.
   end-for
   Fix $A$, $B$, $W$ and approximately solve for $M$.
   Define $F_{\lambda,M_S}(M_Z)=F(M_S)+\langle\nabla F(M_S), M_Z-M_S\rangle+\lambda\|M_Z-M_S\|_F^2/2$.
   for $j$ =1,2,$\cdots$, $h_3$ do
      Set $a_j = (p-1)/p$.
      Compute $M_{S_i} = (1+\alpha_j)M_{Z_i} - \alpha_j M_{Z_{i-1}}$.
      Compute $\nabla_{M_S}F(M_{S_i})$.
      while (true)
         Compute $\widehat{M_S} = M_{S_i} - \nabla_{M_S}F(M_{S_i})/\lambda_i$.
         Compute $[M_{Z_{i+1}}] = \boldsymbol{PSP}(\widehat{M_S})$.
         if $F(M_{Z_{i+1}}) \leq F_{\lambda_i,M_{S_i}}(M_{Z_{i+1}})$, then break;
         else Update $\lambda_i = \lambda_i \times 2$.
         end-if
      end-while
      Update $p = \left(1+\sqrt{1+4p^2}\right)/2$, $\lambda_{i+1}=\lambda_i$.
   end-for
   Fix $A$, $B$, $M$ and approximately solve for $W$.
   Define $F_{\mu,W_S}(W_Z)=F(W_S)+\langle\nabla F(W_S), W_Z-W_S\rangle+\mu\|W_Z-W_S\|_F^2/2$.
   for $j$ =1,2,$\cdots$, $h_4$ do
      Set $a_j = (q-1)/q$.
      Compute $W_{S_i} = (1+\alpha_j)W_{Z_i} - \alpha_j W_{Z_{i-1}}$.
      Compute $\nabla_{W_S}F(W_{S_i})$.
      while (true)
         Compute $W_{Z_{i+1}} = W_{S_i} - \nabla_{W_S}F(W_{S_i})/\lambda_i$.
         if $F(W_{Z_{i+1}}) \leq F_{\mu_i,W_{S_i}}(W_{Z_{i+1}})$, then break;
         else Update $\mu_i = \mu_i \times 2$.
         end-if
      end-while
      Update $q = \left(1+\sqrt{1+4q^2}\right)/2$, $\mu_{i+1}=\mu_i$.
   end-for
end-for
Set $Z^*=[A_{Z_{i+1}}\ B_{Z_{i+1}}\ M_{Z_{i+1}}\ W_{Z_{i+1}}]$.
```

3 Experimental Evaluation

Our experiments are conducted on three publicly available multi-source datasets, namely, UCI Multiple Features (UCI MFeat) [3], Wikipedia [19], and MIR Flickr [9]. The statistics of the datasets are given in Table 2, and brief descriptions of the chosen feature sets in the above-mentioned datasets are listed in Table 3.

Note that all the data are normalized to unit length. Each dataset is randomly separated into a training set and a test set. The training samples account for 80% of each original dataset, and the remaining ones act as the test data.

Table 2. Statistics of the multi-source datasets

Dataset	Total attributes	Total classes	Total samples
UCI MFeat	123	10	2000
Wikipedia	258	10	2866
MIR Flickr	5857	38	25000

Table 3. Brief descriptions of the feature sets

Dataset	Feature set	Total attributes	Total labels	Total instances
UCI MFeat	fou (S_x)	76	10	2000
	zer (S_y)	47	10	2000
Wikipedia	Image (S_x)	128	10	2866
	Text (S_y)	130	10	2866
MIR Flickr	Image (S_x)	3857	38	25000
	Text (S_y)	2000	38	25000

Such a partition of each dataset is repeated five times and the average performance is reported. The 100 outliers are generated from Gaussian noise with the same dimension as normal samples in each source. We mix these outlier data into each dataset. Some key parameters of all the methods in our experiments are tuned using the 5-fold cross-validation based on the AUC (area under the receiver operating characteristic curve) on the training set. Particularly, the LIBSVM classifier serves as the benchmark for the tasks of classification in the experiments.

3.1 Comparison of Multi-view Outlier Detection Methods

The purpose of comparing the proposed MMOD model and multi-view outlier detection methods, such as Li's method [21], Zhao's method [27], Janeja's method [10], and Liu's method [14] is to show the importance of identifying multi-source outliers in a consistent feature-homogeneous space. Due to the attempt of identifying multi-source outliers in different original feature spaces, it is extremely difficulty for the other compared approaches to capture much more complementary information from different sources. It will lead to a low recognition rate for multi-source outliers. To validate this point, we further compare the recognition rate of MMOD with the above-mentioned multi-view outlier detection methods. The parameter settings in the compared methods are the same as in their original literatures.

The proposed MMOD model identifies multi-source outliers in a consistent feature-homogeneous space. As shown in Table 4, MMOD can improve more effectively the recognition rate for multi-source outliers than Li's method, Zhao's method, Janeja's method, and Liu's method. It means that MMOD can capture the information compatibility among different sources more effectively.

Table 4. Comparison of multi-view outlier detection methods

Method	Dataset		
	UCI MFeat	Wikipedia	MIR Flickr
Li's	0.7492	0.7822	0.8053
Zhao's	0.7192	0.7618	0.7891
Janeja's	0.7197	0.7359	0.8096
Liu's	0.8013	0.8296	0.8394
MMOD	**0.8977**	**0.8671**	**0.8824**

3.2 Comparison of Mono-source Outlier Detection Approaches

To evaluate the performance of outlier detection, we compare our method with some representative state-of-the-art mono-source methods such as Li's method [13], Rahmani's method [18], and kNN [6] in three multi-source datasets. Basic metric, Precision (P), is used to evaluate the ability of each algorithm. For Li's method, Rahmani's method, and kNN, we first use CCA [22] to project the multi-source data into a feature-homogeneous space and then apply these methods to retrieve the most likely outlier. For MMOD, we tune the regularization parameters on the set $\{10^i | i = -2, -1, 0, 1, 2\}$. For Li's method and Rahmani's method, the experiment settings follow the original works [13,18], respectively. The parameter k in kNN is selected from the set $\{2 * i + 1 | i = 5, 10, 15, 20, 25\}$.

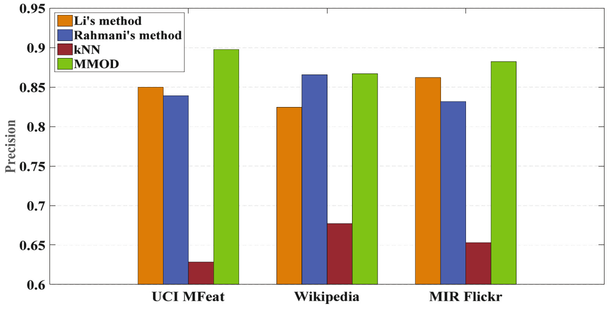

Fig. 3. Comparisons of outlier detection approaches

From Fig. 3, we can see that MMOD achieves significant gains, and can almost detect all the outliers. This observation indicates that MMOD will be more favorable to detect multi-source heterogeneous outliers because of fully taking into account the information compatibility and semantic complementarity among different sources.

3.3 Comparison in Different Outlier Rates

To test the performance of the proposed MMOD in different outlier rates, we further compare the recognition rate of MMOD with other multi-view outlier detection methods such as Li's method [21], Zhao's method [27], Janeja's method [10], and Liu's method [14] in the larger MIR Flickr dataset. We tune the outlier rates on the set $\{10\%, 15\%, 20\%, 25\%\}$.

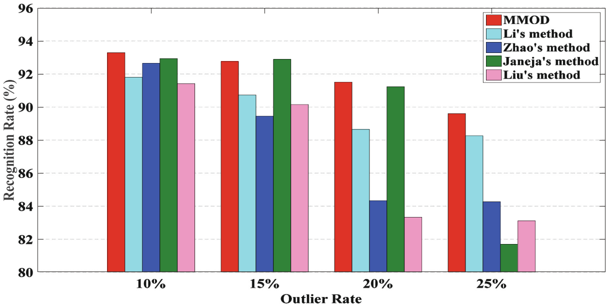

Fig. 4. Comparison in different outlier rates

We can see from Fig. 4 that MMOD is superior to other multi-view outlier detection methods in recognition rate. This observation further confirms that MMOD can effectively identify multi-source outliers. Nevertheless, with the increasing of outlier rate, the performance of MMOD will degrade. Thus, MMOD also has some limitations that it need a certain number of existing samples to identify multi-source outliers.

4 Conclusion

In this paper, we have investigated the heterogeneous outlier detection problem in multi-source learning. We developed a MMOD framework based the consistent representations for multi-source heterogeneous data. Within this framework, Manifold learning is integrated to obtain a shared-representation space, in which the information-correlated representations are close along manifold while the semantic-complementary instances are close in Euclidean distance. Meanwhile, an affine subspace is learned through affine combination of shared representations from different sources in the feature-homogeneous space according to the information compatibility among different sources. Finally, multi-source heterogeneous outliers can be effectively identified in the affine subspace.

Acknowledgments. This work was supported by National Natural Science Foundation of China (No. 61601458, 61602465).

References

1. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. J. Mach. Learn. Res. **6**(3), 1817–1853 (2005)
2. Bertsekas, D.P.: Convex Optimization Theory. Athena Scientific (2009)
3. Breukelen, M.V., Duin, R.P.W., Tax, D.M.J., Hartog, J.E.D.: Handwritten digit recognition by combined classifiers. Kybernetika -Praha- **34**(4), 381–386 (1998)
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. **41**(3), 1–58 (2009)
5. Chen, D., Lv, J., Yi, Z.: A local non-negative pursuit method for intrinsic manifold structure preservation. In: Proceedings of AAAI Conference on Artificial Intelligence, pp. 1745–1751 (2014)
6. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Trans. Inf. Theory **13**(1), 21–27 (2002)
7. Elhamifar, E., Vidal, R.: Sparse subspace clustering: algorithm, theory, and applications. IEEE Trans. Pattern Anal. Mach. Intell. **35**(11), 2765–2781 (2013)
8. Guo, Y., Xiao, M.: Cross language text classification via subspace co-regularized multi-view learning. In: Proceedings of ACM International Conference on Machine Learning, pp. 915–922 (2012)
9. Huiskes, M.J., Lew, M.S.: The MIR Flickr retrieval evaluation. In: Proceedings of ACM International Conference on Multimedia Information Retrieval, pp. 39–43 (2008)
10. Janeja, V., Palanisamy, R.: Multi-domain anomaly detection in spatial datasets. Knowl. Inf. Syst. **36**(3), 749–788 (2013)
11. Knorr, E.M., Ng, R.T.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of International Conference on Very Large Data Bases, pp. 392–403 (1998)
12. Knorr, E.M., Ng, R.T., Tucakov, V.: Distance-based outliers: algorithms and applications. VLDB J. **8**(3), 237–253 (2000)
13. Li, X., Lv, J., Yi, Z.: An efficient representation-based method for boundary point and outlier detection. IEEE Trans. Neural Net. Learn. Syst. **29**(1), 51–62 (2016)
14. Liu, A., Lam, D.: Using consensus clustering for multi-view anomaly detection. In: IEEE Symposium on Security and Privacy Workshops, pp. 117–124 (2012)
15. Nesterov, Y.: Introductory lectures on convex optimization. Appl. Optim. **87**(5), 236 (2004)
16. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
17. Pimentel, M.A.F., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Process. **99**(6), 215–249 (2014)
18. Rahmani, M., Atia, G.: Randomized robust subspace recovery and outlier detection for high dimensional data matrices. IEEE Tran. Signal Process. **65**(6), 1580–1594 (2017)
19. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of ACM International Conference on Multimedia, pp. 251–260 (2010)
20. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
21. Sheng, L., Ming, S., Yun, F.: Multi-view low-rank analysis for outlier detection. In: Proceedings of SIAM International Conference on Data Mining, pp. 748–756 (2015)

22. Sun, L., Ji, S., Ye, J.: Canonical correlation analysis for multilabel classification: a least-squares formulation, extensions, and analysis. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 194–200 (2011)
23. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**(1), 207–244 (2009)
24. Wen, Z., Yin, W.: A feasible method for optimization with orthogonality constraints. Math. Program. **142**(1–2), 397–434 (2013)
25. Zhang, L., et al.: Collaborative multi-view denoising. In: Proceedings of ACM SIGKDD International Conference on Knowledge Discovery Data Mining, pp. 2045–2054 (2016)
26. Zhang, L., Zhao, Y., Zhu, Z., Wei, S., Wu, X.: Mining semantically consistent patterns for cross-view data. IEEE Trans. Knowl. Data Eng. **26**(11), 2745–2758 (2014)
27. Zhao, H., Fu, Y.: Dual-regularized multi-view outlier detection. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 4077–4083 (2015)