Partial Multi-View Outlier Detection Based On Collective Learning

Ajinkya Rawankar, Pranav Verma, Amrit Kataria, Pranjal Patidar

1. Introduction

In real world, data is collected from various domains and data from each domain is called a view. A video can be represented as visual and audio information , an image can be represented by color, shape and other features and a webpage can be represented by images, words, and URLs on the page.

Multi-view outlier detection is detecting the anomalies data from multi-view dataset. Anomalies in multi-view data are instances that have inconsistent features across multiple views. The task is called horizontal anomaly detection, or multi-view anomaly detection. It is more challenging due to the complicated distribution and organisation of data. Note that each view is a set of features and each view is may have different feature dimensions.

Collective learning is the ability to share information so efficiently that the ideas of individuals can be stored within the collective memory of communities and can accumulate through generations.

2. Real life application

In real-world applications, each view may suffer from some missing samples, which results in partial objects some of such applications are :

- Intrusion detection
- Social media
- Health care

3. Dataset Used

Handwritten datasets from two popular datasets i.e. **MNIST** and **USPS**. MNIST consist of dataset contains 70000 digit images with the size of 28 × 28 and USPS dataset contains 9298 digit images with the size of 16 × 16. The same two datasets can be regarded as two different views, sincethey were collected under different scenarios.

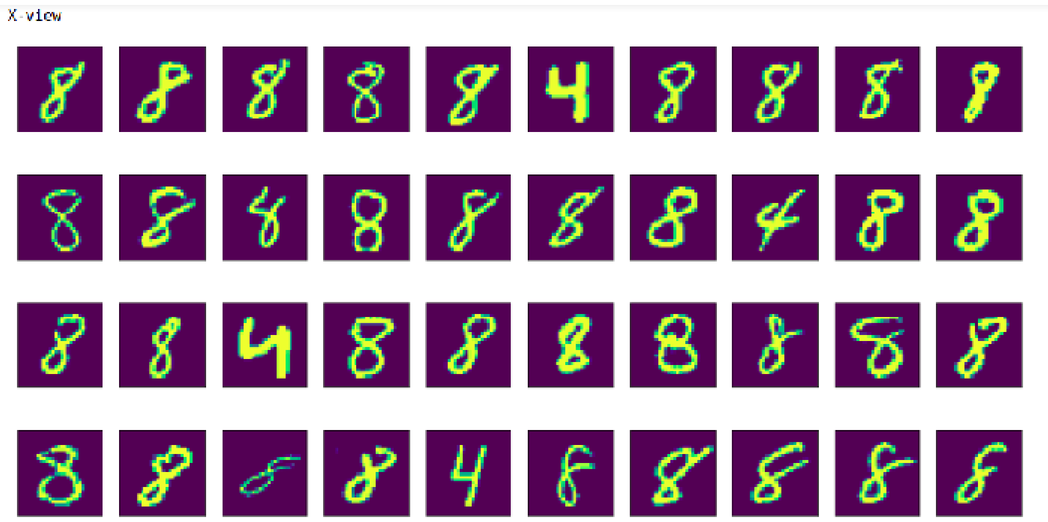
4. Data Preprocessing

From the datasets, select 50 images per digit from each dataset randomly. So there are 500 samples in each view such that i^{th} sample of I view is same as i^{th} sample of in II view.

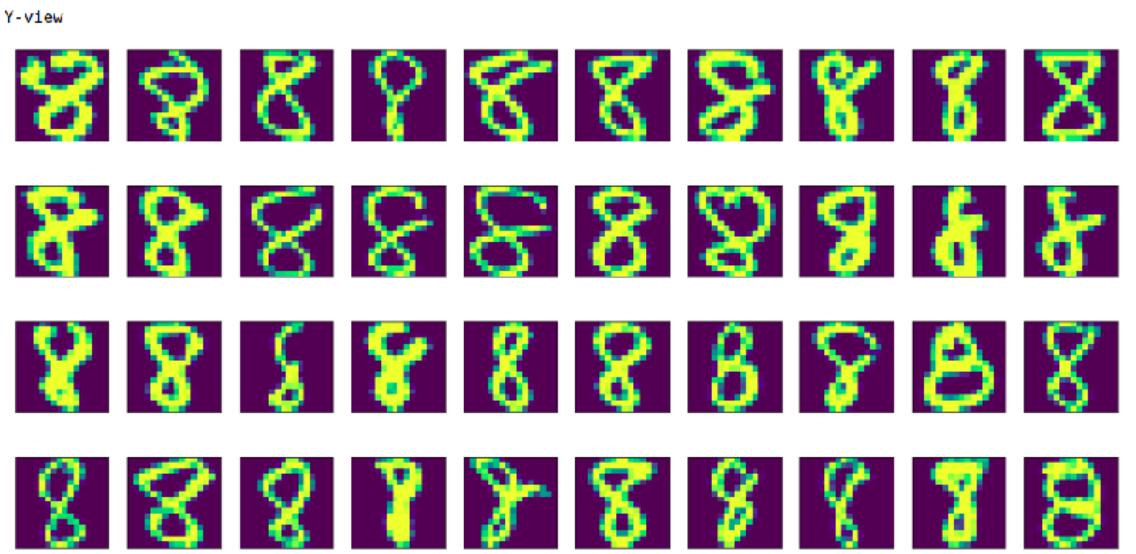
But these samples are complete, so we delete some % of data from (let say 20%) each view such that i^{th} data from I view and j^{th} data from II view such that i != j .

Next we have to create outliers in data swap 10% of data with one another catagory of data. Finally we will get partial multi view data with outliers.

Data for Digit 8 from MNIST (with outliers)



Data for Digit 8 from USPS (without outliers)



5. HSIC

The paper implements kernel based statistical independence measure called the HSIC (Hilbert Schmidt Independence criterion). The HSIC measure utilises a cross covariance operator 'C', and its empirical estimation is given as :

$$\text{HSIC} = \frac{1}{n^2} tr(KHLH)$$

$$\text{where, } H = I_n - \frac{1}{n} 11^T$$

The independence measure tells us how independent two random variables from each other. This forms the basis for the collective learning implemented in the paper.

Here, linear kernels, $L = XX^T$ and $K = YY^T$, have been used as the kernel.

6. KNN

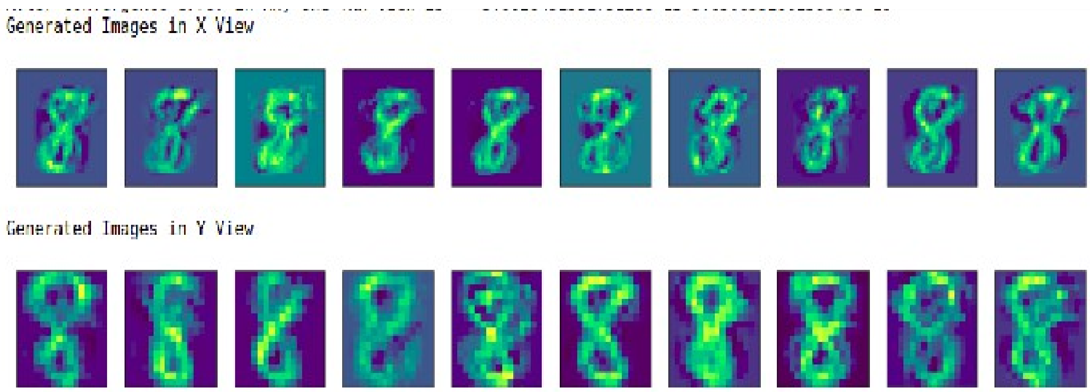
k-NN is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms. It treats all features equally.

In this Project we used Knn with some improvements , i.e. we don't have labels and for each data point we are finding k data points which are nearest neighbours of that data point.

7. Algorithm Implementation

First we generate the missing data. For this, we take the average of each category and replace missing data with average values.

Now we have to do collective learning. For that we use Hilbert-Schmidt Independence Criterion (HSIC). We begin by initialising the confidence matrix $C = 11^T$. Then calculate the P matrix as $P = \text{diag}(C)HXX^TH$ and similarly Q for Y and update X^{ny} and Y^{nx} in X and Y respectively, and run this until the algorithm converges. We apply KNN to compute the Similarity matrix W^x and W^y such that W[i][j]=1 if **j** is in list of nearest neighbours of **i**. From this calculating the Delta as Hw^xHW^y and generate a vector s (score) i.e. diagonal of Delta. After scaling score vector update confidence matrix $C=SS^T$ and iterate until T exhausts.



8. Hyperparameters

Hyperparameter	Value
K in knn	7
Threshold	0.2
POR	0.2
T	100

9. Challenges faced

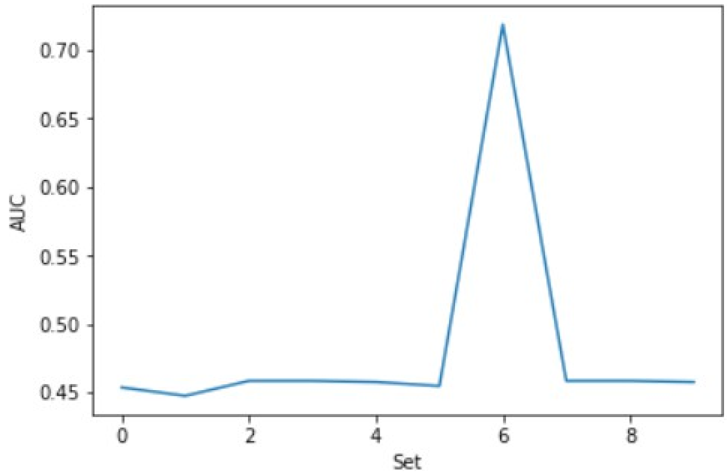
- Understanding of significance of HSIC (Hilbert-Schmidt Independence Criteria)
- Verifying the result
- Parameter tuning :
  - Threshold selection
  - No. of self guided iterations
  - No. of iterations for convergence of the missing data for each view

10. Result And Analysis

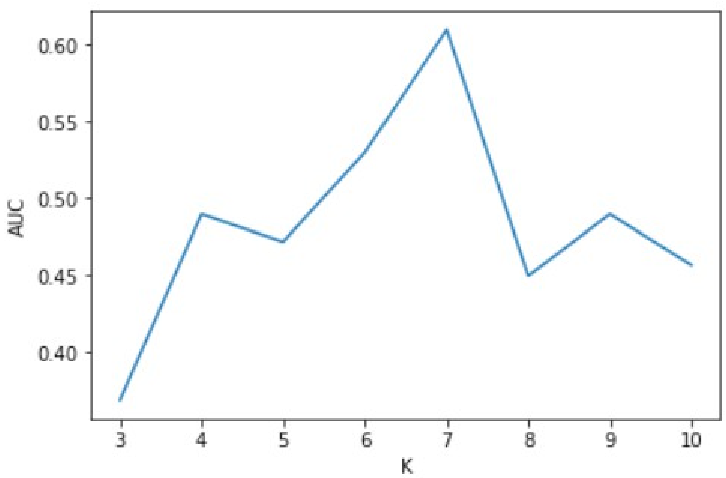
Outlier detection for partial multi-view data is evaluated by AUC, i.e., the area under Receiver Operating Characteristic (ROC) curve.

We executed the algorithm on different k values and checked on each digit class and found the following trend :

Plot of set(digit class) vs AUC :



Plot of k-value vs AUC :



11.Contributions

1. Data preprocessing : Amrit Kataria
2. Data generation using HSIC implementation : Pranav Verma
3. KNN and similarity matrix generation : Ajinkya Rawankar
4. Confidence matrix generation and outlier indicator vector : Pranjal Patidar
5. Report : All
6. Presentation : All

12. References

[1] Masashi Sugiyama,Makoto Yamada, "On Kernel Parameter Selection in Hilbert-Schmidt Independence Criterion" IEICE Transactions on Information and Systems, 1 vol.E95-D, no.10, pp.2564–2567, 2012.

[2] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Scholkopf, Alexander J. Smola "A Kernel Statistical Test of Independence".

[3] Jun Guo, Wenwu Zhu "Partial Multi-View Outlier Detection Based on Collective Learning" The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).

[4] https://en.wikipedia.org/wiki/Independence_(probability_theory)"

[5] https://en.wikipedia.org/wiki/Feature_selection