SHIFT15M: Fashion-specific dataset for set-to-set matching with several distribution shifts

Masanari Kimura  
ZOZO Research  
amasanari.kimura@zozo.com

Takuma Nakamura  
ZOZO Research  
takuma.nakamura@zozo.com

Yuki Saito  
ZOZO Research  
yuki.saito@zozo.com

Abstract

Set-to-set matching is the problem of matching two different sets of items based on some criteria. Especially when each item in the set is high-dimensional, such as an image, set-to-set matching is treated as one of the applied problems to be solved by utilizing neural networks. Most machine learning-based set-to-set matching generally assumes that the training and test data follow the same distribution. However, such assumptions are often violated in real-world machine learning problems. In this paper, we propose SHIFT15M, a dataset that can be used to properly evaluate set-to-set matching models in situations where the distribution of data changes between training and testing. Some benchmark experiments show that the performance of naive methods drops due to the effects of the distribution shift. In addition, we provide software to handle SHIFT15M dataset in a very simple way: https://github.com/st-tech/zozo-shift15m. The URL for the software will appear after this manuscript is published.

1. Introduction

One of the key problems for fashion data analysis is set-to-set matching [1, 2, 23]. For example, we can consider a task that measures the degree of completion of an outfit by matching sets of clothing items (i.e., for two sets $A = \{\text{hat, shirt, skirt}\}$ and $B = \{\text{jacket, shoes}\}$, the matching score of $A$ and $B$ corresponds to the goodness of the outfit $A \cup B$). To solve this, we need to investigate neural networks that handle sets [13, 15, 26, 28, 35, 36, 39, 41].

Another common phenomenon in the domain of fashion is trend change. These phenomena are observed at various scales, ranging from annual trend changes such as fashionable colors to seasonal trend changes such as summer to winter clothing. In the field of machine learning, such an assumption can be defined as a distribution shift (or dataset shift) [11, 18, 19, 22, 24, 27, 37].

Many machine learning problem settings assume that training and test data are independent and identically distributed (i.i.d.). We assume that training examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n_{tr}}$ are independently and identically distributed (i.i.d.) according to some fixed but unknown distribution $p_{tr}(x, y)$, which can be decomposed into the marginal distribution and the conditional probability distribution, i.e., $p_{tr}(x, y) = p_{tr}(x)p_{tr}(y|x)$. We also denote the test examples by $\{(x^{(i)}_{te}, y^{(i)}_{te})\}_{i=1}^{n_{te}}$ drawn from a test distribution $p_{te}(x, y) = p_{te}(x)p_{te}(y|x)$.
Figure 4. Several sample images from SHIFT15M dataset.

Table 1. Statistics on the SHIFT15M dataset.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#sets</td>
<td>2,555,147</td>
<td>1,423</td>
<td>4,813</td>
<td>131,611</td>
<td>466,583</td>
<td>730,443</td>
<td>617,844</td>
<td>299,502</td>
<td>137,510</td>
<td>92,944</td>
<td>59,412</td>
<td>13,062</td>
</tr>
<tr>
<td>#items</td>
<td>15,218,721</td>
<td>8,327</td>
<td>29,140</td>
<td>756,532</td>
<td>2,644,564</td>
<td>4,305,802</td>
<td>3,731,864</td>
<td>1,853,647</td>
<td>855,036</td>
<td>576,022</td>
<td>373,549</td>
<td>84,238</td>
</tr>
<tr>
<td>mean set size</td>
<td>6.03</td>
<td>5.85</td>
<td>6.05</td>
<td>5.74</td>
<td>5.66</td>
<td>5.89</td>
<td>6.04</td>
<td>6.18</td>
<td>6.21</td>
<td>6.19</td>
<td>6.28</td>
<td>6.44</td>
</tr>
<tr>
<td>median set size</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>5.00</td>
<td>5.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td>mean #likes</td>
<td>26.98</td>
<td>0.94</td>
<td>2.00</td>
<td>15.74</td>
<td>16.84</td>
<td>23.24</td>
<td>37.37</td>
<td>35.67</td>
<td>32.41</td>
<td>24.89</td>
<td>21.34</td>
<td>16.01</td>
</tr>
<tr>
<td>median #likes</td>
<td>9.00</td>
<td>0.00</td>
<td>1.00</td>
<td>8.00</td>
<td>6.00</td>
<td>13.00</td>
<td>18.00</td>
<td>23.00</td>
<td>19.00</td>
<td>17.00</td>
<td>12.00</td>
<td>12.00</td>
</tr>
<tr>
<td>#unique users</td>
<td>193,574</td>
<td>289</td>
<td>571</td>
<td>16,922</td>
<td>52,283</td>
<td>80,290</td>
<td>49,441</td>
<td>18,854</td>
<td>7,511</td>
<td>4,442</td>
<td>2,739</td>
<td>853</td>
</tr>
</tbody>
</table>

Definition 1.1. (Covariate shift [25]) We consider that the two distributions $p_{tr}(x, y)$ and $p_{te}(x, y)$ satisfy the covariate shift assumption if the following conditions hold:

$$p_{tr}(x) \neq p_{te}(x), \quad p(y|x) = p_{tr}(y|x) = p_{te}(y|x).$$

Definition 1.2. (Target shift [40]) We consider that the two distributions $p_{tr}(x, y)$ and $p_{te}(x, y)$ satisfy the target shift assumption if the following conditions hold:

$$p_{tr}(y) \neq p_{te}(y), \quad p(x|y) = p_{tr}(x|y) = p_{te}(x|y).$$

To address these problems, we provide SHIFT15M, a real-world dataset that can handle the above two problem settings, that is, the set-to-set matching dataset with distribution shift. Our SHIFT15M dataset is built on data accumulated over the past 10 years in our fashion SNS. In this SNS, users could post combinations of their clothing items and other users could bookmark them as favorites. The data accumulated by this service, which has been in operation for a decade from 2010 to 2020, is very useful for dealing with distribution shifts in the fashion sector. Figure 2 shows an overview of the SHIFT15M dataset. Each column is a set of posted fashion items, with information such as the user who posted, the date of publication, and the price of each item. We hope that our SHIFT dataset will encourage research on set-to-set matching tasks under the distribution shift.

1.1. Contribution

Our contributions are summarized as follows:

- We propose SHIFT15M, a fashion-specific dataset that can properly evaluate models for set-to-set matching under the distribution shift assumptions. SHIFT15M also enables the performance evaluation of the model under various magnitudes of dataset shifts by switching the magnitude. Figure 4 shows several sample images from the SHIFT15M dataset.

- We provide open-source software to handle the SHIFT15M dataset in a very simple way. Figure 3 shows the minimum sample code of our software;

- We propose first-step benchmark methods for set-to-set matching under distribution shift, numerical experiments show the usefulness of these methods.

2. Statistics on the SHIFT15M dataset

In this section, we present some statistics for our SHIFT15M dataset. First, Table 1 shows the overview of statistics on the SHIFT15M dataset. Since our fashion SNS was launched in 2010, the number of users and posts gradually increased from 2010, reaching a peak around 2014~2015, and slowly decreasing until 2020, the year the service was terminated. Also, the number of items in a set tends to increase over the years, indicating that users tend to construct outfits with more and more items.

Top panel of Figure 5 shows the trend of price for items included in SHIFT15M. Bottom panel: trend of number of likes for posted sets.

- We provide open-source software to handle the SHIFT15M dataset in a very simple way. Figure 3 shows the minimum sample code of our software;

- We propose first-step benchmark methods for set-to-set matching under distribution shift, numerical experiments show the usefulness of these methods.
Figure 6. Trends of the number of posted sets by year.

Figure 7. Covariate shift of image features.

Figure 8. Category classification results under the covariate shift.

included in the SHIFT15M dataset. It can be seen that the fashion items posted by users are becoming more expensive every year. Bottom panel of Figure 5 shows the trend of number of likes for posted sets.

Figure 6 plots the trend of the number of posted sets by year. This figure shows that our fashion SNS, the source of the SHIFT15M dataset, was most active around 2014-2015.

Finally, we confirm the covariate shift of the image features included in SHIFT15M. If covariate shift assumption 1.1 holds, we should be able to construct a classifier \( f: \mathbf{x} \mapsto y = \{0, 1\} \), where \( \mathbf{x} \) is the image feature of the item and \( y \) is the binary classification output for two years. Figure 7 shows the experimental results. The results show that classification between distant years (e.g., acc. of 2010 vs. 2020 is 0.85) is easier, while classification between close years (e.g., acc. of 2010 vs. 2011 is 0.62) is more difficult, indicating a gradual shift in image features. Figure 8 also shows the experimental results of item categorization when the training and test data were generated from different years. This figure shows that the closer the years of the training and test data are, the higher the classification accuracy.

3. Benchmarks

In this section, we introduce several numerical experiments on the SHIFT15M dataset.

3.1. Importance weighted set-to-set matching

As the benchmark strategy for the distribution shift adaptation on the set-to-set matching, we propose importance weighted set-to-set matching which is based on IWERM.

**Definition 3.1.** (Importance weighted ERM [25]) Importance Weighted Empirical Risk Minimization (IWERM) uses the density ratio \( p_{te}(\mathbf{x})/p_{tr}(\mathbf{x}) \) as the weighting function:

\[
\hat{h} = \arg \min_{h \in \mathcal{H}} \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} p_{te}(\mathbf{x}_{i}^{tr}) f(h(\mathbf{x}_{i}^{tr}), y_{i}^{tr}).
\]

Adopting the density ratio as the weighting function, as in Definition 3.1, leads to the following statistically important property.

**Theorem 3.1.** (Consistency of IWERM [25]) If we set \( w(\mathbf{x}) = p_{te}(\mathbf{x})/p_{tr}(\mathbf{x}) \) as the weighting function, the empirical error computed by the weighted ERM is consistent estimator of the expected error in the test distribution.

Using the above ideas, we propose a novel covariate shift adaptation method for set-to-set matching. Let \( \mathcal{L}(\mathcal{V}, \mathcal{W}, f) \) be the K-pair-set loss [23] function for the set matching, which is defined as follows:

\[
\mathcal{L}(\mathcal{V}, \mathcal{W}, f) = -\frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \delta_{ij} \log \frac{\exp(f(\mathcal{V}_{i}, \mathcal{W}_{j})))}{\sum_{k=1}^{K} \exp(f(\mathcal{V}_{i}, \mathcal{W}_{k}))},
\]

where \( \delta \) is Kronecker’s delta, and we can modify \( \mathcal{L}(\mathcal{V}, \mathcal{W}, f) \) as follows:

\[
\mathcal{L}_{w}(\mathcal{V}, \mathcal{W}, f) = -\frac{1}{K} \sum_{i=1}^{K} \sum_{j=1}^{K} \delta_{ij} \Gamma_{ij}^{p} \log \frac{\Gamma_{ij}^{f}}{\sum_{k=1}^{K} \Gamma_{i,k}^{f}},
\]

where \( \Gamma_{ij}^{p} = \exp(p(\text{test} | \mathcal{V}_{i} \cup \mathcal{W}_{j})) \) and \( \Gamma_{ij}^{f} = \exp(f(\mathcal{V}_{i}, \mathcal{W}_{j})). \) This modification can be regarded as a weighting based on the probability that the pair is included in the test set. Here, we propose two weighting strategies:

\[
\text{max-IW} : p(\text{test} | \mathcal{V}_{i} \cup \mathcal{W}_{j}) = \max_{x \in \mathcal{V}_{i} \cup \mathcal{W}_{j}} w(x),
\]

\[
\text{mean-IW} : p(\text{test} | \mathcal{V}_{i} \cup \mathcal{W}_{j}) = \frac{1}{|\mathcal{V}_{i} \cup \mathcal{W}_{j}|} \sum_{x \in \mathcal{V}_{i} \cup \mathcal{W}_{j}} w(x),
\]

where \( w(x) \) is the weighting function. Next, we approximate \( w(x) \) by using unlabeled data from both \( p_{tr} \) and \( p_{te} \). In IWERM, the squared error can be decomposed as follows:

\[
\mathbb{E}_{p_{tr}}[\Delta^{2}] = \mathbb{E}_{p_{tr}}[\hat{w}(x)\Delta^{2}] + \mathbb{E}_{p_{tr}}[(w(x) - \hat{w}(x))\Delta^{2}],
\]
This indicates that the weighting function is approximated by the function \( g(x) \). 

\[
\Delta^2 = \| f(x) - y \|^2 \quad \text{and} \quad \hat{w}(x) \text{ is the approximator of the weighting function } w(x). 
\]

Second term is bounded as

\[
\mathbb{E}_{p_{tr}} \left[ (w(x) - \hat{w}(x)) \Delta^2 \right] 
\leq \frac{1}{2} \left( \mathbb{E}_{p_{tr}} \left[ \Delta^2 \right] + \mathbb{E}_{p_{tr}} \left[ (w(x) - \hat{w}(x))^2 \right] \right)
\]  \tag{3}

Let \( s \) be the indicator of the distributions, where \( s = 1 \) corresponds to the train distribution and \( s = 0 \) corresponds to the test distribution, and we assume that \( p(s) = 0.5 \). Then, we also assume that

\[
p(x|s) = \begin{cases} 
    p_{tr}(x) & (s = 1), \\
    p_{te}(x) & (s = 0). 
\end{cases}
\]  \tag{4}

Then, we have \( w(x) = \mathbb{E}_{p(x|s=0)} \frac{E(x|s=1)p(s=1|x)}{1 + \hat{w}(x)}. \) Let \( g(x) \) be the optimal source discriminator which identifies whether \( x \) is generated \( p_{tr} \) or \( p_{te} \). Then, we can write as \( g(x) = p(s = 1|x) = \frac{1}{\hat{w}(x) + 1} \). Suppose that the density ratio \( p_{te}(x)/p_{tr}(x) \) is bounded by \( \beta > 0 \), we have \( \frac{1}{1 + \hat{w}(x)} \leq g(x) \leq 1 \) for all \( x \). From the unlabeled data generated from \( p_{tr} \) and \( p_{te} \), we can learn the estimator \( \hat{g} \) of \( g \). Then, we can write the weight estimation term as

\[
\mathbb{E}_{p_{tr}} \left[ (w(x) - \hat{w}(x))^2 \right] = \mathbb{E}_{p_{tr}} \left[ \left( \frac{g(x) - \hat{g}(x)}{\hat{g}(x)} \right)^2 \right]
\]

\[
\leq (1 + \beta)^4 \mathbb{E}_{p_{tr}} \left[ (g(x) - \hat{g}(x))^2 \right] 
\]

\[
= (1 + \beta)^4 \mathbb{E}_{p_{te}} \left[ (g(x) - \hat{g}(x))^2 \frac{p_{te}(x)}{p_{te}(x)} \right] 
\]

\[
\leq 2(1 + \beta)^4 \mathbb{E}_{p_{te}} \left[ (g(x) - \hat{g}(x))^2 \right] 
\]

\[
= 2(1 + \beta)^4 \left\{ \mathbb{E}_{p_{te}} \left[ (s - g(x))^2 \right] - \mathbb{E}_{p_{te}} \left[ (g(x) - \hat{g}(x))^2 \right] \right\}.
\]

This indicates that the weighting function is approximated by the function \( g(x) \).

### 3.2. Results of numerical experiments

We introduce benchmark results for a set-to-set matching under the covariate shift. The model architecture is same as the previous work \([23]\), which is based on the architecture of Transformer \([15, 20, 33]\). Our task can be considered an extended version of a standard task, Fill-In-The-Blank \([8]\), which requires us to select an item that best extends an outfit from among four candidates. Because selecting a set corresponds to filling multiple blanks, we consider the set matching problem as Fill-In-The-N-Blank \([23]\). To construct the correct pair of sets to matched, we randomly halve the given outfit \( O \) into two non-empty proper subsets \( V \) and \( W \), as follows: \( O \rightarrow \{V, W\} \), where \( V \cap W = \emptyset \).

Tables 2 and 3 show the experimental results of the Fill-In-The-N-Blank with four and eight candidates. In these experiments, data from 2013 are used as training data and data from 2013–2017 are used as test data. ERM refers to empirical risk minimization \([6, 31, 32]\), which assumes that \( p_{tr}(x) = p_{te}(x) \). From these results, we can see that the covariate shift adaptive set-to-set matching methods can achieve better performances than the ordinal ERM.

### 4. Related works and conclusion

Several distribution shift datasets exist for general classification and regression tasks where the input is a vector. WILDS \([14]\) is the collection of benchmark datasets \([3–5, 7, 9, 12, 17, 21, 29, 38]\) under the distribution shift, including histopathological images, satellite images or sequence of source code tokens. PACS \([16]\) and Office-Home \([34]\) adopt the image style to differentiate distributions, and VLCS \([10]\) takes data collected independently from four sources as environments. Also, DomainNet \([42]\) extends PACS to a far larger scale, consisting of more domains and categories.

We believe that our SHIFT15M is a very useful dataset for evaluating the still underdeveloped task of set-to-set matching under natural distribution shifts.
References


generalization: A survey.  


