Modal Consistency based Pre-Trained Multi-Model Reuse*

Yang Yang De-Chuan Zhan Xiang-Yu Guo Yuan Jiang
National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China
{yangy, zhandc, guoxy, jiangy}@lamda.nju.edu.cn

Abstract

Multi-Model Reuse is one of the prominent problems in Learnware [Zhou, 2016] framework, while the main issue of Multi-Model Reuse lies in the final prediction acquisition from the responses of multiple pre-trained models. Different from multi-classifiers ensemble, there are only pre-trained models rather than the whole training sets provided in Multi-Model Reuse configuration. This configuration is closer to the real applications where the reliability of each model cannot be evaluated properly. In this paper, aiming at the lack of evaluation on reliability, the potential consistency spread on different modalities is utilized. With the consistency of pre-trained models on different modalities, we propose a Pre-trained Multi-Model Reuse approach (PM^{2}R) with multi-modal data, which realizes the reusability of multiple models. PM^{2}R can combine pre-trained multi-models efficiently without re-training, and consequently no more training data storage is required. We describe the more realistic Multi-Model Reuse setting comprehensively in our paper, and point out the differences among this setting, classifier ensemble and later fusion on multi-modal learning. Experiments on synthetic and real-world datasets validate the effectiveness of PM^{2}R when it is compared with state-of-the-art ensemble/multi-modal learning methods under this more realistic setting.

1 Introduction

Machine learning techniques have achieved great success in many applications. Nowadays, many machine learning models are integrated as a part of functional software. Similar to the developments of software engineering research, machine learning researchers pay more attentions on the reuse of pre-trained learners in real applications [Zhou, 2016; Pan and Yang, 2010].

*This work was supported by NSFC (61632004, 61673201), Huawei Fund (YBN2017030027) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

Model reuse is with different levels: transfer learning techniques reuse models by obtaining middle level representations [Wei *et al.*, 2016; Long and Wang, 2015; Isele *et al.*, 2016] for cross domain flexibility of models. However, it is notable that in transfer learning both source domain and target domain data are required during the training phase, i.e., transfer learning is a technical engineer oriented reuse.

Nevertheless, learnware describes a substantially different paradigm of model reuse: pre-trained models are ready with specifications which are sufficient for users. In general, users can fetch a group of similar models for their task. Note that there are no source domain training examples in this kind of model reuse, and pre-trained learners cannot be modified in their model essentially. This type of reuse paradigm is more practical in real cases since most users can only understand the specifications instead of managing the skills of training/refining the model parameters. In addition, although Zhou [2016] reserves the interface of retaining model in learnware, users cannot afford re-training on the target domain in practice for lacking of their own techniques and experiences, computational consumptions, and more important, the lack of labeled data in target domain. Thus, the learnware paradigm tends to propose a re-training free and general user oriented model reuse.

How to *reuse multi-model predictions* becomes the most important problem in this case. Since ideally, one can directly find a pre-trained model with the exact same specification as the target domain requirement. However, due to the differences between the environments of pre-training and the user's target domain, it is more common that one cannot find a "perfect" pre-trained model but a set of analogous models. We denote this problem as *Multi-Model Reuse* (MMR) problem in the context. MMR problem is similar to the scene when a customer chooses the same type of goods in the supermarket. Ideally, if the customer is well experienced and with demands precisely described, he can directly buy the appointed goods. However, a more real case is that he may take more time on hesitating for equal confidence assignments to the set of items. In MMR problem, users can also use some ensemble techniques, e.g., the majority voting, for the confidence assigned equally when there is no prior knowledge provided. Moreover, in ensemble learning, we can use weighted voting for better performance since model-wise confidence can be obtained with evaluation on the ground truth. Nev-

ertheless, in MMR problem, labeled data are absent, and the confidence re-weighting cannot be obtained directly. Similar phenomenon occurs in the market example, it is also ridiculous to ask the feelings before customers buy items.

Nevertheless, one in a supermarket can eventually make the decision by reading the specifications, comparing the "detail" and "consistency" of those descriptions. The "consistency" in production specification can be described from multiple aspects (modalities). In multi-modal learning, researchers have explored the consistency of predictions from different modalities for semi-supervised learning [Blum and Mitchell, 1998; Wang *et al.*, 2016], clustering [Li *et al.*, 2014; Xu *et al.*, 2015], theoretical analysis [Wang and Zhou, 2013] etc., and these show that the consistency among multiple modalities can be very informative in the situation where precision evaluation cannot be applied.

In this paper, we propose the Pre-trained Multi-Model Reuse method, PM^{2}R, to tackle the MMR problem under the multi-modal learning scenario. Different from ensemble methods, no training data and validation instances are provided in the MMR problem, thus PM^{2}R is a re-training free strategy. Comparing to the multi-modal late fusion methods, the MMR emphasizes user oriented reuse and requires the pre-trained models fixed, therefore, PM^{2}R is built with no modification on pre-trained models. PM^{2}R can be implemented from various ways, yet in this paper, a belief propagation style updating is demonstrated and the effectiveness of this PM^{2}R implementation is shown in the experiment.

The rest of this paper starts from introduction of related work. Then we propose our approach, followed by experiments and conclusion.

2 Related Work

This work focuses on reuse of multiple models and is related to multi-classifier systems, transfer learning and multi-modal learning. In this section, we first review these topics and figure out the main differences between these existing work and our new settings eventually.

Multi-classifier systems or know as ensemble learning are proposed to increase generalization abilities of single type of base classifiers. Breiman [1996] proposed Bagging, which trains a component model from several training sets generated from the original training set. Schapire [1990] proposed Boosting, which generates a series of component neural networks whose training sets are determined by the performance of former ones. Zhou *et al.* [2002] analyzed the ensemble learning, pointed that with a portion of selected base learners, the ensemble learner can be with even higher generalization ability than using all base learners and consequently proposed the GASEN approach which selects a portion of neural networks based on the evolved weights to make up the ensemble; [Ueda, 2000] optimized linear weights to combine component predictions based on statistical pattern recognition theory. Stacking [Wolpert, 1992] is a special kind of ensemble which can be regarded as a complicated vote scheme generated by a second level classifier. The majority voting can be used for our MMR problem for lacking of prior knowledge on pre-trained models, however, an obvious gap between MMR problem and ensemble techniques lie in the availability of labeled training data and the existence of modifiable models. In the MMR setting, no labeled data can be revisited since there are rarely labeled training examples stored in learnware repositories, and no apparent adjustments can be made by inexperienced users.

Transfer learning concentrates on solving the prediction tasks on target domains. In recent, many transfer learning approaches are proposed for expanding the scope of learning applications, e.g., Tan *et al.* [2017] proposed a selective learning algorithm to solve the distant domain transfer learning problem, Kandemir [2015] performed knowledge transfer by projecting the target data onto the source domain and linearly combining its representations on the source and target domain manifolds. However, similar issues in ensemble learning are also appeared on solving MMR with transfer learning: it requires both source and target labeled data for transfer learning, while in MMR setting, only unlabeled data can be gathered by model reuse users (denoted as user domain data without any confusion).

Without labeled examples in user domain, model evaluation cannot be precisely established in MMR setting, and consequently, it may be hard to re-weight the confidences of multiple models with classification precision. Applications with complicated object descriptions are usually brought forward the problem of multi-modal learning. In multi-modal learning, different modalities can provide uncorrelated predictions and consistency becomes an important evaluation than precision. Blum and Mitchell [1998] trained two classifiers separately on two modalities and then uses them to label unlabeled instances for each other, Wang and Zhou [2013] presented the theoretical analysis on co-training when neither modality is sufficient. Multi-modal learning approaches can be categorized into 3 types, i.e., pre-fusion, subspace, late-fusion, and late fusion is more closed to MMR problem, e.g., Wang *et al.* [2013] proposed the WNH to integrate all modal features and learn the corresponding weights for every modality, Ye *et al.* [2015] proposed RANC in a hybrid fusion manner. However, both in WNH and RANC, the classifiers of different modalities are adjusted in different iterations during the training phase, which is not agreed with the MMR setting.

There have been some model reuse approaches, e.g., FMR [Yang *et al.*, 2017] which tries to integrate the discriminative ability of fixed models into one deep network, and eventually can be applied in various applications. However, in our MMR settings, a different barrier of model reuse is tackled, i.e., weighting or selection of proper multiple models for reuse is mainly focused.

We propose the PM^{2}R approach for classifying new instances directly from a collection of responses from different pre-trained models with various modalities. Note that PM^{2}R can be implemented from various ways, yet a belief propagation style updating is demonstrated in this paper.

3 Proposed Method

3.1 Notations

In this paper, without any loss of generality, suppose there are N instances on the user's side, which are denoted by

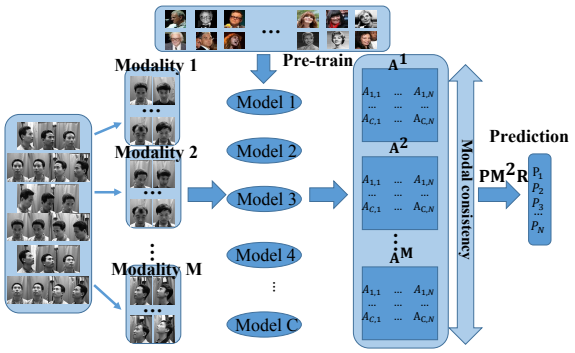


Figure 1: The overall flowchart. Test instances are with various modalities, while the pre-trained multi-models are from other datasets (the upper side); the predictions gathered are formed into matrices $\{A^1, A^2, \cdots, A^M\}$, the final predictions are obtained by PM^{2}R with the consistency among different modalities considered. Note that PM^{2}R is built with no modification on pre-trained models.

$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$. Each instance has $d-$dimensional inputs as raw features, i.e., $\mathbf{x}_i = [x_{i_1}, x_{i_2}, \cdots, x_{i_d}] \in \mathbb{R}^d$. Meanwhile, in multi-modal learning, instance space can be denoted as, at least M parts without any overlap, $v = \{v_1, v_2, \cdots, v_M\}$, where $v_m \in \mathbb{R}^{d_m}$ is raw features from $j-$th modality, $d = d_1 + d_2 + \cdots + d_M$. Therefore, in multi-modal settings, instance $\mathbf{x}_i$ can be further denoted as $(\mathbf{x}_{i,v_1}, \mathbf{x}_{i,v_2}, \cdots, \mathbf{x}_{i,v_M})$, where $\mathbf{x}_{i,v_m}$ denotes the raw feature representation of the m-th modality of i-th instance. In this paper, those multiple modalities are homogeneous ones, e.g., face photographs of the same person with different expressions or from different poses. Suppose there are C pre-trained models in "learnware market" chosen and denoted as $\{f_1, f_2, \cdots, f_C\}$. The corresponding prediction of user's side instances on a concerned modality is $A^m_{i,j} = f_j(\mathbf{x}_{i,v_m})$, and $A^m_{i,j} \in \{-1, +1\}$.

3.2 The Pre-trained Multi-Model Reuse Method

PM^{2}R focuses on the MMR problem under the homogeneous multi-modal learning scenario. Different from ensemble methods, there are no training data or validation set provided in the MMR settings. Meanwhile, the MMR emphasizes user oriented reuse and requires the pre-trained models fixed, therefore, PM^{2}R is built with no modification on pre-trained models. In practice, as a re-training free strategy, PM^{2}R only has the corresponding predictions of user's side as inputs. More specially, for the $m-$th modality, we can denote the prediction values of all the instances as $A^m \in \mathbb{R}^{N \times C}$, where

$$A^m = \left\{A^m_{i,j}, \quad i \in \{1, 2, \cdots, N\}, \ \ j \in \{1, 2, \cdots, C\}\right\}.$$

That is, for each modality, C classifiers are invoked for predicting and returns N predictions for all instances. Assignment bipartite graph $G^m(X^m, V, E^m)$ is further used for representing the responses between instances and models in detail, where the superscript m represents the m-th modality, X^m is the set of instances on the m-th modality, V is the set of pre-trained models, and $(i^m, j) \in E^m$ means the j-th classifier makes prediction on instance i on modality m. Eventually, there are M assignment bi-graphs $\{G^1, G^2, \cdots, G^M\}$ due to the multi-modal setting. The overall flowchart is shown in Fig. 1.

To solve the MMR problem under this setting, we claim that the modal consistency should be considered and PM^{2}R can be implemented in various ways. Yet in this paper, a self-consistent weight propagation strategy is used for passing the weights between the two sides of bi-graph, and a message passing style algorithm operating two types of messages are proposed for seeking the equilibrium state of the weight propagation system. In detail, two types of messages are used, i.e., the instance messages $\{z_{i^m \to j}\}$, $(i^m, j) \in E^m$ capturing how likely instance i being predicted as positive on the $m-$th modality, and the model messages $\{y_{j \to i^m}\}$ capturing how reliable the j-th model is.

Implementation by Considering Multi-Modal as One

First, we can concatenate the prediction values of each modality, denoted as $A = [A^1, A^2, \cdots, A^M] \in \mathbb{R}^{N \times (CM)}$. Besides, the set of edge assignment between the pre-trained model j and predict instance i^m of different modalities can be concatenated as $E = E^1 \cup E^2 \cdots \cup E^M$ as well. Specially, the instance messages can be denoted as $\{z_{i \to j}\}$, $(i, j) \in E$, and the model messages can be represented as $\{y_{j \to i}\}$. In each round, all the messages are updated as follows:

$$\begin{aligned} z_{i \to j} &= \sum_{j' \neq j} A_{i,j'} y_{j' \to i} \\ y_{j \to i} &= \sum_{i' \neq i} A_{i',j} z_{i' \to j}, \end{aligned} \tag{1}$$

where the first is generally the weighted majority voting according to how reliable every model is, and the second is updating the reliability according to how many times the model agreed. It is notable that the inherent updating of model messages utilizes all the instance with multi-modal information. This approach is denoted as the PM^{2}R$_{\text{One}}$ in the following text. The pseudo code of PM^{2}R$_{\text{One}}$ is given in Algorithm 1.

Implementation by Considering Consistency Explicitly

Besides considering multi-modal as one, a more reasonable approach lies considering the consistency between different modalities. The PM^{2}R$_{\text{One}}$ method implicitly uses multi-modal information, while we also implement PM^{2}R by utilizing the consistency explicitly between different modalities in iterations. In PM^{2}R all messages are updated separately according to the consistency index function $\mathbb{I}(z_{i^m \to j} \sum_m z_{i^m \to j} \geq 0)$. $I(\text{stat.}) = 1$ iff. the statement stat. is true, which figures out whether modal consistency meets. In this version, A^m and E^m are independent for different modalities. In detail, after passing the message from model to instance, the instance messages of each modality, i.e., $z_{i^1 \to j}, \cdots, z_{i^M \to j}$, are re-consisted and updated according to following equations:

$$z_{i^m \to j} = \begin{cases} z_{i^m \to j}, & (z_{i^m \to j} \sum_m z_{i^m \to j} \geq 0), \\ \dfrac{\sum_m z_{i^m \to j} \mathbb{I}(z_{i^m \to j} \sum_m z_{i^m \to j} \geq 0)}{\sum_m \mathbb{I}(z_{i^m \to j} \sum_m z_{i^m \to j} \geq 0)}, & \text{otherwise}, \end{cases} \tag{2}$$

The pseudo code is described in Algorithm 2.

Algorithm 1 The pseudo code of PM^{2}R$_{\text{One}}$

Input:
$A \in \mathbb{R}^{N\times(CM)}$ is outputs of C classifiers on M modalities
$E = E^1 \cup E^2 \cdots \cup E^M$, is the union of all linkages between multi-modal features and pre-trained models
K is the maxIter
Output:
$\{p_1, p_2, \cdots, p_N\}$: the final predictions for all N instances

for all $(i,j) \in E$ **do**
  Initialize model messages $y^0_{j\to i}$ with random values under Normal distribution
end for
for $k = 1, 2, \cdots, K$ **do**
  for all $(i,j) \in E$ **do**
    $z^k_{i\to j} = \sum_{j'\neq j} A_{i,j'} y^{(k-1)}_{j'\to i}$
  end for
  for all $(i,j) \in E$ **do**
    $y^k_{j\to i} = \sum_{i'\neq i} A_{i',j} z^k_{i'\to j}$
  end for
end for
for all i **do**
  $p_i = \text{sign}\left(\sum_j A_{i,j} y^K_{j\to i}\right)$,
end for

Algorithm 2 The pseudo code of PM^{2}R

Input:
$\{A^1, A^2, \cdots, A^M\} \in \mathbb{R}^{N\times C}$ are outputs of C classifiers on M modalities seperately
$\{E^1, E^2, \cdots, E^M\}$ are the linkages between different modal instances and pre-trained models
K is the maxIter
Output:
$\{p_1, p_2, \cdots, p_N\}$: the final predictions for all N instances

Initialize $y^0_{j\to i^m}$, $m \in \{1, 2, \cdots, M\}$ as steps 1-3 in Algorithm 1
for $k = 1, 2, \cdots, K$ **do**
  Update $z^k_{i^m\to j}$, $m \in \{1, 2, \cdots, M\}$ as steps 5-7 in Algorithm 1
  for all $(i^m, j) \in E^m$ **do**
    Update $z^k_{i^m\to j}$ as in Eq. 2
  end for
  Update $y^k_{j\to i^m}$, $m \in \{1, 2, \cdots, M\}$ as steps 8-10 in Algorithm 1
end for
$p_i = \text{sign}\left(\sum_m \sum_j A^m_{i,j} y^K_{j\to i^m}\right)$, $i \in \{1, 2, \cdots, N\}$

4 Experiment

In the MMR problem, PM^{2}R$_{\text{One}}$/PM^{2}R can classify new instances directly from a collection of predictions provided by a bundle of different pre-trained models with various modalities. Besides, PM^{2}R$_{\text{One}}$/PM^{2}R are built without modification on pre-trained models.

In this section, we will provide the empirical investigations and performances of PM^{2}R$_{\text{One}}$/PM^{2}R from the users' side of model reuse. In particular, investigations on both synthetic dataset and real-world tasks, i.e., gender classification, are studied. The real-world task is to predict the gender of people with different modalities, i.e., in different expressions, poses, etc. There are 3 different subsets in gender prediction task.

Some of the ensemble methods without re-training can be used in MMR problem, thus, PM^{2}R$_{\text{One}}$/PM^{2}R are compared to the widely used ensemble method, i.e., Majority Voting (denoted as MV). For PM^{2}R$_{\text{One}}$/PM^{2}R are related to late fusion style multi-modal learning approaches, WNH, RLF, RANC are also compared in our experiments. The test sets are drawn from the users' side data with bootstrap, and repeated for 30 times for each task. For both synthetic and real world task, the average accuracies and standard deviations are recorded. In detail, the compared methods are listed as:

- **MV**: classifies instances from different modalities with the collected predictions provided by different pre-trained models via majority voting, i.e., suppose there are M modalities and C classifiers, then the majority voting is performed on $M \times C$ prediction values;
- **WNH**: combines all prediction values from different modalities together and then uses $l_{2,1}$-norm to regularize the modality selection process and finally gives the prediction [Wang *et al.*, 2013];
- **RLF**: minimizes the rank of indicator matrix to maximize the modal consistency and fuse the predicted confidence scores of multiple models [Ye *et al.*, 2012]. Note that RLF can only provide the final prediction rather than predictions on single modality;
- **RANC**: truncated low-rank regularized approach for fusing the prediction values, modal consistency is considered through truncated nuclear norm regularization. Final fusing steps in [Ye *et al.*, 2012] are used;
- **BPE**: belief propagation style method. In implementation, belief propagation is invoked on every modality, and then majority vote across different modalities for final predictions [Yedidia *et al.*, 2003].

4.1 Synthetic Data

Synthetic data are generated according to [Khetan and Oh, 2016]. For every modality, the prediction confidence of instances i is parameterized by $q_i \in [0, 1]$ which represents how likely the instance is positive. More specifically, when $j-$th pre-trained model predicts the $i-$th instance on $m-$th modality as positive, we assign the instance as positive with probability q_i. Hence, when q_i is close to 0.5, the instances on the concerned modality can be regarded as confusing and is difficult to correctly classify. Meanwhile, the j-th pre-trained model is parameterized by $p_j \in [0, 1]$, which represents how trustable the model is. The closer p_j approaches one, the more trustable this model is. p_j closed to 0.5 indicates the model providing a random guess. With the q_i and p_j defined above, we can get the expected prediction values of $A^m_{i,j}$ as:

$$\mathbb{E}(A^m_{i,j}) = \begin{cases} 1 & w.p. \quad q_i p_j + \bar{q}_i \bar{p}_j, \\ -1 & w.p. \quad \bar{q}_i p_j + q_i \bar{p}_j. \end{cases} \tag{3}$$

Table 1: The accuracy (avg.± std.) of compared methods on synthetic data. The best classification performance is bolded.

	Single Modality					Final
MV	.754±.012	.756±.009	.757±.013	.765±.011	.771±.010	.901±.008
WNH	.751±.012	.753±.010	.755±.013	.762±.012	.768±.010	.899±.008
RANC	.779±.013	.778±.008	.770±.011	.781±.012	**.817±.011**	.936±.011
BPE	.806±.007	.813±.005	.806±.007	.815±.008	.810±.007	.950±.006
RLF	-	-	-	-	-	.935±.005
PM^{2}R$_{\text{One}}$	.808±.007	**.815±.006**	.810±.008	**.817±.008**	.811±.008	.952±.007
PM^{2}R	**.808±.006**	.813±.006	**.811±.005**	.814±.008	.812±.007	**.967±.004**

Table 2: The accuracy (avg.± std.) of compared methods on *-Exp. data. The best classification performance is bolded.

Acc.-Exp.	Single Modality					Final
MV	.835±.017	.843±.017	.827±.014	.822±.009	.835±.012	.856±.013
WNH	.816±.018	.861±.017	.821±.015	.813±.007	.832±.013	.875±.014
RANC	**.899±.001**	.811±.000	**.894±.002**	**.894±.000**	**.894±.000**	**.894±.001**
BPE	.816±.015	.840±.015	.797±.008	.816±.011	.813±.011	.827±.012
RLF	-	-	-	-	-	572.±.016
PM^{2}R$_{\text{One}}$	.817±.015	.840±.014	.797±.008	.816±.011	.819±.011	.827±.012
PM^{2}R	.819±.012	**.864±.010**	.798±.008	.848±.012	.820±.010	.888±.007

Pos.-Exp.	Single Modality					Final
MV	.864±.010	.859±.011	.888±.010	.840±.009	.856±.012	.912±.009
WNH	.817±.010	.789±.011	.832±.010	.789±.009	.813±.012	.888±.010
RANC	**.936±.001**	**.936±.000**	**.936±.000**	**.936±.000**	**.936±.000**	.936±.000
BPE	.832±.011	.840±.010	.843±.013	.824±.008	.843±.009	.861±.007
RLF	-	-	-	-	-	.519±.019
PM^{2}R$_{\text{One}}$	.832±.012	.843±.010	.832±.013	.824±.006	.840±.007	.861±.006
PM^{2}R	.869±.013	.864±.012	.875±.013	.851±.007	.875±.008	**.957±.007**

WIKI-Exp.	Single Modality					Final
MV	.904±.011	.922±.013	.896±.012	.885±.011	.875±.009	.941±.007
WNH	.803±.011	.843±.012	.856±.012	.816±.011	.795±.009	.918±.007
RANC	.872±.001	.872±.000	.872±.000	.872±.002	.872±.000	.872±.002
BPE	.795±.010	.827±.011	.835±.009	.829±.007	.801±.009	.825±.007
RLF	-	-	-	-	-	.561±.029
PM^{2}R$_{\text{One}}$	.806±.009	.826±.012	.827±.008	.832±.007	.803±.008	.825±.006
PM^{2}R	**.913±.008**	**.915±.010**	**.875±.011**	**.894±.009**	**.878±.010**	**.975±.006**

In synthetic experiments, we can directly generate the prediction values $A^m_{i,j}$. In practice, we generated the values for 1000 instances with 5 different modalities. The synthetic label for each instance is sampled i.i.d. from a Bernoulli distribution with parameter 0.5. Meanwhile, the difficulty q_i is sampled from a Beta distribution with mean 0.75 (for positive instances) or 0.25 (for negative instances), and the variance is configured as 0.04. We re-sample the task difficulty for each modality to make the q_i varies between different modalities. Then we construct 70 pre-trained models whose reliability p_j are sampled from a Beta distribution with mean 0.6 and variance 0.01. Pre-trained models are sampled only once for all modalities. Finally, the prediction values of different modalities are calculated based on Eq. 3, and denoted as $\{A^1, A^2, \cdots, A^M\}$. PM^{2}R$_{\text{One}}$/PM^{2}R are invoked for the prediction of model reuse, and prediction accuracies are recorded in Table 1. From the table, it reveals that for both single modality and final prediction, PM^{2}R$_{\text{One}}$/PM^{2}R almost always achieve the best classification performance on mean accuracy comparing to all other methods.

4.2 Gender Classification

In the real-world gender classification task, one of the most widely used multi-modal datasets is used, i.e., CAS-PEAL [Gao *et al.*, 2008] is constructed by Chinese Academy of Sciences (CAS). The CAS-PEAL is with large-scale face images from different sources of variations, specifically, the variations include poses, expressions, accessories, etc. The whole dataset contains 99,594 images from 1040 individuals (595 males and 445 females). The CAS-PEAL is naturally divided into 4 subsets with different variations, i.e., personal poses (denoted as Pos.) set containing 1038 instances with different poses/modalities, expressions (denoted as Exp.) set containing 376 instances with different expressions/modalities, accessories (denoted as Acc.) set including 434 instances with different accessories/modalities. For each subset in Pos., there are 9 different shooting slopes, i.e., there are 9 cameras spaced equally in a horizontal semicircular shelf to simultaneously capture images across different angles in one shot. Besides, each subset also contains two group of visual directions, i.e., all people are asked to look up and down to capture 18 images in another two shots, therefore, we can divide the Pos. set into 21 modalities. We also divide Exp. set into 5 modalities, and similarly, 6 modalities for Acc. set (photoed with 3 different glasses or with 3 different caps). More detailed descriptions on subsets categorization can be found in [Gao *et al.*, 2008].

40 pre-trained models are generated on Pos. set, i.e., we extract traditional BOW features, Fisher vectors features, LBP features, HOG features from Pos., then train 12 random forest models with different number of trees, 24 support vector models with different kernel methods or costs using these features respectively, besides, 4 deep models are also included, which are trained with the raw features. Note that we can treat the pre-trained models as a bundle of black-boxes in whole from the "learnware market" side, while from the aspect of users' side, all 40 models are pre-trained and unmodifiable. Users of learnware models can only input raw pictures from source datasets, e.g., Exp. set, consequently the black-boxes of pre-trained models will output the prediction values, and eventually the final predictions are obtained with PM^{2}R$_{\text{One}}$/PM^{2}R together with compared methods. For facilitating the notations, this setting is denoted as Pos.-Exp, we also have investigations of Pos.-Acc. and Acc.-Exp..

To demonstrate the generalization ability of PM^{2}R$_{\text{One}}$ /PM^{2}R. We conduct more experiments with external data sources. We utilize the WIKI [Rothe *et al.*, 2015] dataset, which is also a face dataset with the same input size, for models pre-training, and predict with Pos., Exp., Acc. sets separately. WIKI is one of the largest datasets of face images publicly available with gender and age labels. It contains 62,328 profiled images from pages of people from Wikipedia with their profiles date of birth, name, gender and all images associated with that person. Therefore, we also have following settings as WIKI-Pos., WIKI-Exp. and WIKI-Acc..

From Table 2 to Table 4, it can be observed that although the models on single modality don't provide satisfactory performance in partial subsets due to the dis-matching of users side tasks, PM^{2}R can almost always achieve the best performance on the final precision, except for Acc.-Exp. set.

Table 3: The accuracy (avg.± std.) of compared methods on WIKI-Pos. data. The best classification performance is bolded.

WIKI-Pos.	Partial Single Modality												Final
MV	.841±.010	.805±.009	.908±.016	.799±.009	.822±.009	.857±.006	.855±.017	.825±.007	.846±.005	.798±.014	.884±.010	.813±.012	.920±.006
WNH	.820±.009	.802±.011	.876±.014	.777±.008	.783±.008	.810±.006	.819±.015	.815±.005	.811±.004	.777±.010	.812±.009	.814±.015	.951±.004
RANC	.916±.007	**.964±.012**	**.994±.012**	**.969±.010**	**.947±.014**	**.945±.006**	**.981±.011**	**.964±.014**	**.964±.007**	**.921±.013**	**.989±.012**	**.947±.003**	.953±.012
BPE	.838±.010	.806±.014	.908±.017	.822±.005	.817±.012	.832±.011	.847±.017	.828±.015	.842±.008	.805±.014	.845±.010	.807±.015	.911±.008
RLF	-	-	-	-	-	-	-	-	-	-	-	-	.633±.025
PM^2R_{one}	.837±.010	.806±.007	.902±.015	.821±.009	.819±.013	.836±.013	.846±.011	.826±.014	.842±.003	.821±.011	.853±.017	.807±.010	.919±.012
PM^2R	**.947±.005**	.865±.017	.921±.004	.883±.013	.862±.010	.900±.011	.916±.011	.894±.014	.885±.011	.867±.013	.890±.011	.881±.003	**.971±.008**

Table 4: The accuracy (avg.± std.) of compared methods on *-Acc. data. The best classification performance is bolded.

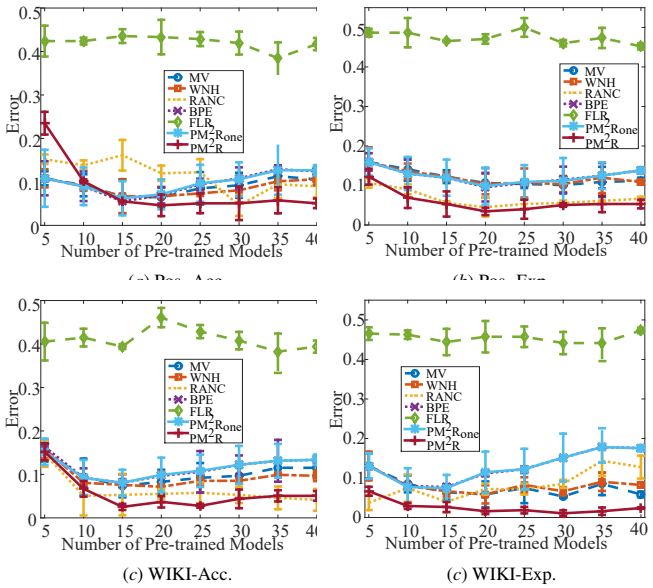
Pos.-Acc.	Single Modality					Final
MV	.818±.019	.843±.010	.845±.013	.811±.007	.799±.018	.899±.014
WNH	.774±.019	.811±.010	.822±.013	.779±.007	.758±.018	.898±.014
RANC	.763±.010	.847±.019	**.951±.014**	**.868±.013**	.696±.009	.868±.011
BPE	.809±.010	.827±.020	.831±.016	.797±.006	.783±.014	.870±.012
RLF	-	-	-	-	-	.638±.025
PM^2R_{one}	.809±.008	.827±.015	.831±.018	.797±.007	.786±.011	.873±.010
PM^2R	**.818±.012**	**.864±.017**	.854±.017	.813±.004	**.827±.015**	**.926±.010**

WIKI-Acc.	Single Modality					Final
MV	.815±.011	.822±.016	.841±.009	.809±.012	.813±.017	.884±.011
WNH	.786±.011	.807±.016	.813±.010	.772±.013	.802±.018	.907±.012
RANC	.722±.020	.778±.019	.822±.024	.823±.023	**.847±.012**	.945±.018
BPE	.795±.011	.818±.013	.831±.011	.786±.013	.809±.010	.866±.008
RLF	-	-	-	-	-	.631±.025
PM^2R_{one}	.795±.008	.818±.015	.832±.018	.786±.007	.809±.011	.866±.010
PM^2R	**.816±.015**	**.830±.011**	**.844±.008**	**.824±.013**	.823±.012	**.947±.014**

This phenomenon clearly reveals the effectiveness of considering modal consistency in model reuse. Without considering the modal consistency explicitly, PM^2R_{one} becomes inferior. The performance of RANC is superior on single modality in most case, however, the final prediction performance of RANC is not as well as PM^2R. This may be simply because RANC is a multi-modal learning method and do not solve the problem of striding over the gap between the source (original models) and target (users side) domains.

4.3 Influences of Number of Models for Reuse

In order to explore the influence on the number of pre-trained models, more experiments are conducted. In this section, the number of modality in each investigation is fixed, while the number of pre-trained models varies in $\{5, 10, \cdots, 40\}$. The average errors and standard deviations are recorded in Fig. 2. Due to the page limits, we only list 4 datasets for verification, i.e., Pos.-Acc., Pos.-Exp., WIKI-Acc. and WIKI-Exp.. From these figures, it clearly shows that PM^2R achieves the best performance when the number of models is larger than 15. However, without explicitly considering in the modal consistency, the performance of PM^2R_{one} is inferior to PM^2R. Besides, we can also find that PM^2R achieves a stable performance fast, and the errors of our model reuse strategies decrease faster than compared methods, as the number of pre-trained models increases. It is notable that with the number of models increasing, the error would not decrease without limits, especially for PM^2R_{one}, the error may increase after



(c) WIKI-Acc. (c) WIKI-Exp.

Figure 2: Influences of number of models on gender tasks

the number of models is over a threshold. This may indicate it is wise to select a proper number of models rather than all feasible models in learnware repositories.

5 Conclusion

Model Reuse has attracted many attentions in machine learning communities recently. Multi-Model Reuse (MMR) problem is one of the challenge issues in Learnware. Aiming at the problem of the final prediction acquisition from multiple prediction values in MMR, we follow the learnware principle and propose a multi-model reuse method under multi-modal scenario: PM^2R. Different from multi-classifiers ensemble, there are only pre-trained models rather than the whole training sets provided in this setting. The difficulty of MMR lies lacking the evaluation on reliability for pre-trained models, PM^2R solves this by utilizing the potential consistency on different modalities. Experiments on synthetic and real-world datasets validate the effectiveness of PM^2R under this realistic setting. It is interesting to expand PM^2R style approaches to multi-class scenario and provide theoretical analysis on both generalization abilities and convergence.

References

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.

[Breiman, 1996] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[Gao *et al.*, 2008] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. *IEEE Transactions on System Man and Cybernetics*, 38(1):149–161, 2008.

[Isele *et al.*, 2016] David Isele, Mohammad Rostami, and Eric Eaton. Using Task Features for Zero-Shot Knowledge Transfer in Lifelong Learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 1620–1626, New York, NY, 2016.

[Kandemir, 2015] Melih Kandemir. Asymmetric Transfer Learning with Deep Gaussian Processes. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 730–738, Lille, France, 2015.

[Khetan and Oh, 2016] Ashish Khetan and Sewoong Oh. Reliable Crowdsourcing under the Generalized Dawid-Skene Model. *arXiv:1602.03481*, 2016.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial Multi-View Clustering. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 1968–1974, Quebec City, Canada, 2014.

[Long and Wang, 2015] Mingsheng Long and Jianmin Wang. Learning Transferable Features with Deep Adaptation Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 97–105, Lille, France, 2015.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[Rothe *et al.*, 2015] Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep EXpectation of Apparent Age from a Single Image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 252–257, Santiago, Chile, 2015.

[Schapire, 1990] Robert E. Schapire. The Strength of Weak Learnability. *Machine Learning*, 5(2):197–22, 1990.

[Tan *et al.*, 2017] Ben Tan, Yu Zhang, Sinno Jialin Pan, and Qiang Yang. Distant Domain Transfer Learning. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, San Francisco, CA, 2017.

[Ueda, 2000] Naonori Ueda. Optimal Linear Combination of Neural Networks for Improving Classification Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):207–215, 2000.

[Wang and Zhou, 2013] Wei Wang and Zhi-Hua Zhou. Co-training with Insufficient Views. In *Proceedings of the 5th Asian Conference on Machine Learning*, pages 467–482, Canberra, Australia, 2013.

[Wang *et al.*, 2013] Hua Wang, Feiping Nie, and Heng Huang. Multi-View Clustering and Feature Learning via Structured Sparsity. In *Proceedings of the 30th International Conference on Machine Learning*, pages 352–360, Atlanta, GA, 2013.

[Wang *et al.*, 2016] Yang Wang, Wenjie Zhang, Lin Wu, Xuemin Lin, Meng Fang, and Shirui Pan. Iterative Views Agreement: An Iterative Low-Rank Based Structured Optimization Method to Multi-View Spectral Clustering. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 2153–2159, New York, NY, 2016.

[Wei *et al.*, 2016] Ying Wei, Yin Zhu, Cane Wing ki Leung, Yangqiu Song, and Qiang Yang. Instilling Social to Physical: Co-Regularized Heterogeneous Transfer Learning. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pages 1338–1344, Phoenix, Arizona, 2016.

[Wolpert, 1992] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-View Self-Paced Learning for Clustering. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 3974–3980, Buenos Aires, Argentina, 2015.

[Yang *et al.*, 2017] Yang Yang, De-Chuan Zhan, Ying Fan, Yuan Jiang, and Zhi-Hua Zhou. Deep Learning for Fixed Model Reuse. In *Proceedings of the 31th AAAI Conference on Artificial Intelligence*, pages 1751–1757, Phoenix, Arizona, 2017.

[Ye *et al.*, 2012] Guangnan Ye, Dong Liu, I-Hong Jhuo, and Shih-Fu Chang. Robust Late Fusion with Rank Minimization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3021–3028, Providence, RI, 2012.

[Ye *et al.*, 2015] Han-Jia Ye, De-Chuan Zhan, Yuan Miao, Yuan Jiang, and Zhi-Hua Zhou. Rank Consistency based Multi-View Learning: A Privacy-Preserving Approach. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 991–1000, Melbourne, Australia, 2015.

[Yedidia *et al.*, 2003] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.

[Zhou *et al.*, 2002] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang. Ensembling neural networks: Many could be better than all. *Artificial intelligence*, 137(1-2):239–263, 2002.

[Zhou, 2016] Zhi-Hua Zhou. Learnware: On the Future of Machine Learning. *Frontiers of Computer Science*, 10(4):589–590, 2016.