Supplementary Materials

Learning to Predict With Unavailable Features: an End-to-End Approach via Knowledge Transferring

Chun-Chen Lin * Li-Wei Chang † Chun-Pai Yang ‡ Shou-De Lin §

0.1 Experiment Setups We consider six real-world datasets from UCI Machine Learning Repository [1], Avila, Adult, Hepmass, Eye, Letter and Spam, to quantitatively evaluate the performance of our proposed methods in comparisons with baselines and state-of-the-art imputation methods. Table 1 in our supplementary materials shows the properties of the datasets. In particular, we evaluate the performance using several missing rates on features including 20%, 40% and 60%. Note that the original data of Avila, Adult, Hepmass, Eye, Letter and Spam are fully observed. We randomly drop features and repeat the experiments. The results come from the average of 20 different random droppings.

0.2 Discussions with hyper-parameter analyses Here we provide some in-depth analysis of the proposed model.

Weight-decay Mechanism We analyze the proposed weight-decay mechanism by setting different decay steps. Remind that the model weights for unavailable features are gradually decayed to zero within decay steps. The number of decay steps is inversely proportional to the decay rate. We show the results in Fig. 1 of different decay steps. The performance shows that the decay ratio is not critically sensitive, though tuning it up can still improve the performance.

Teacher-Student Framework Here are two hyper-parameters, the temperature τ and the weight $\lambda \in [0, 1]$ between the distillation loss and the student loss in our teacher-student framework. Empirically, when the student model is very small comparing to the teacher model, lower temperatures work better. Since our teacher model and student model share the same model architecture, we set τ to default value of 1. λ affects how much the knowledge is transferred from the teacher to the student. That is, if the λ is set to be large, the student will consider the information from the teacher more than the information from the ground truth data and vice versa. Note that if λ is set to 0, it is essentially the same as *Drop* method which only utilizes available features during training. We investigate the impact of teacher on student by setting different λ. As shown in Fig. 2, the best results are achieved at λ above 0.5 on the average, implying that the knowledge from the teacher plays an important role when missingness occurs.

GAIN and MisGAN As described in the experiment setup, we only miss those unavailable features and try four different missing probabilities in the training phase to see how GAIN and MisGAN perform in our scenario. As shown in Fig. 3 and Fig. 4, it can be observed that the best performance is achieved when the missing probability is high. We conclude that the higher of *p_miss* and the lower of *obs_prob*, which corresponds to the case of 100% missing probability in the scenario of unavailable features in prediction phase, results in GAIN and MisGAN achieve better performance.

References

[1] M. Lichman, *UCI machine learning repository*, 2013. URL http://archive.ics.uci.edu/ml.

*Dept. of Computer Science and Information Engineering, National Taiwan University. r07922006@ntu.edu.tw

†Dept. of Computer Science and Information Engineering, National Taiwan University. r08922041@ntu.edu.tw

‡Dept. of Computer Science and Information Engineering, National Taiwan University. d05922017@ntu.edu.tw

§Dept. of Computer Science and Information Engineering, National Taiwan University. sdlin@csie.ntu.edu.tw

Table 1: Datasets for model evaluation.

	Avila	**Adult**	**Hepmass**	**Eye**	**Letter**	**Spam**
Instances	20,867	48,842	10,500,000	14,980	20,000	4,601
Features	10	14	27	15	16	57

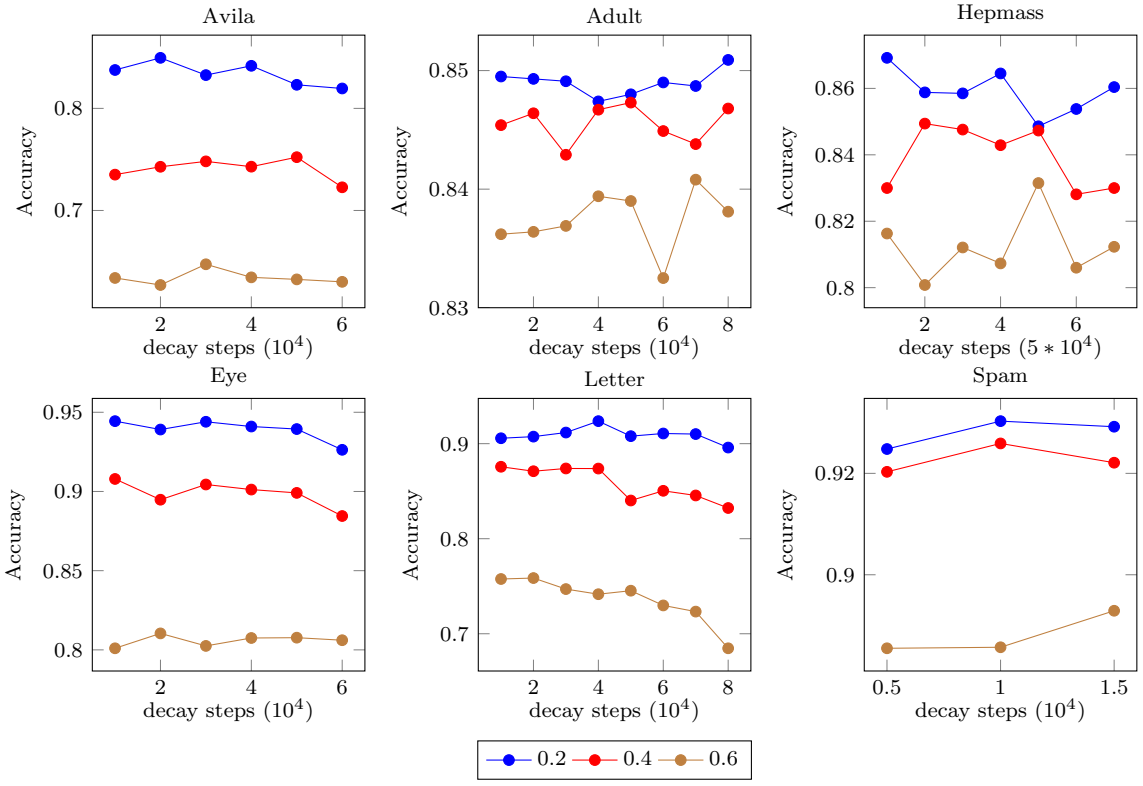

Figure 1: Analysis of weight decaying mechanism with different decaying steps on five datasets with missing ratio 0.2, 0.4 and 0.6.

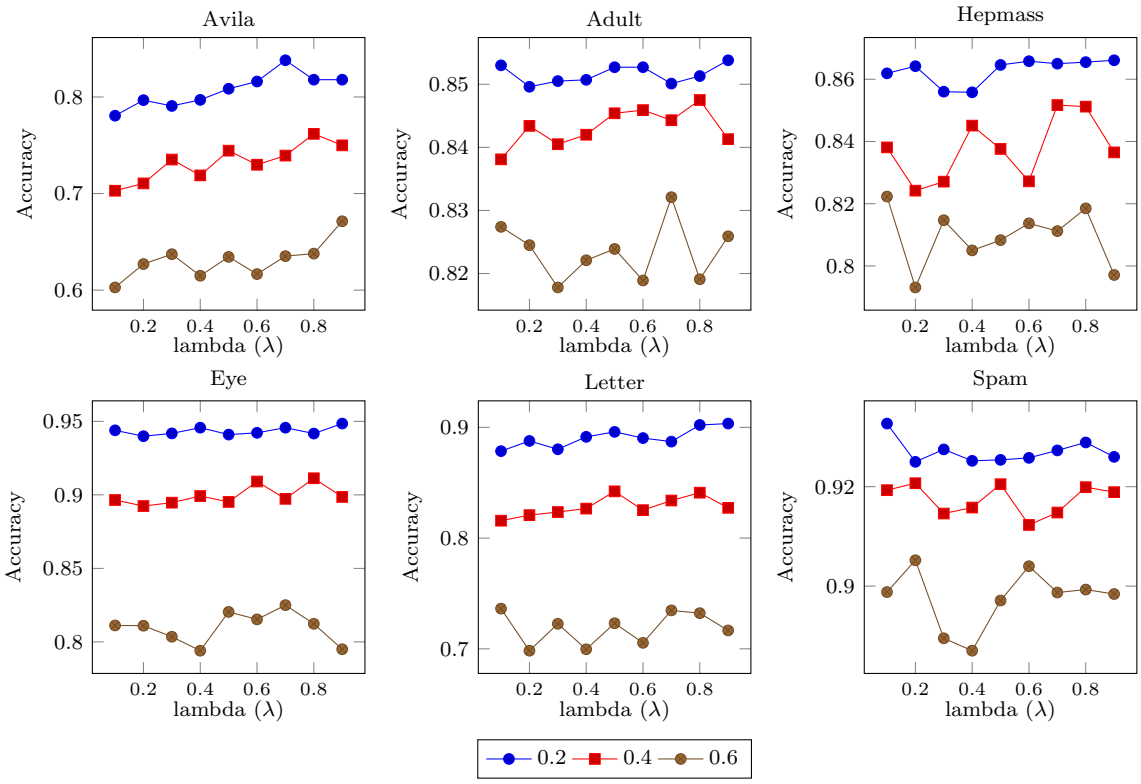

Figure 2: Analysis of teacher-student framework with different λ on five datasets with missing ratio 0.2, 0.4 and 0.6.

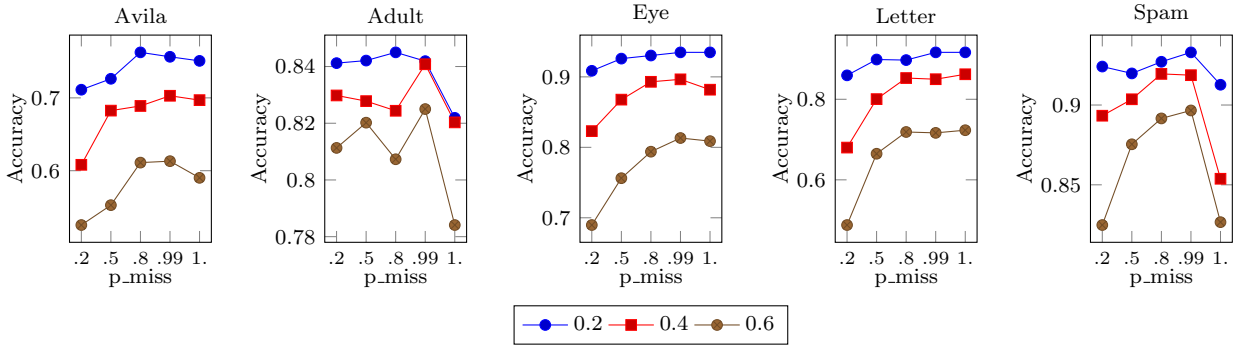

Figure 3: Analysis of GAIN with different missing probabilities on five datasets with missing ratio 0.2, 0.4 and 0.6.

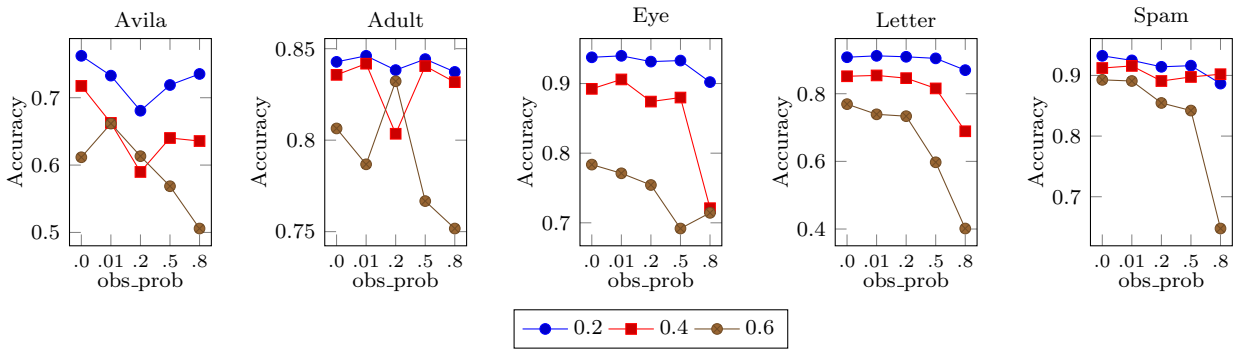

Figure 4: Analysis of MisGAN with different observing probabilities on five datasets with missing ratio 0.2, 0.4 and 0.6.