Robustness to Missing Features using Hierarchical Clustering with Split Neural Networks

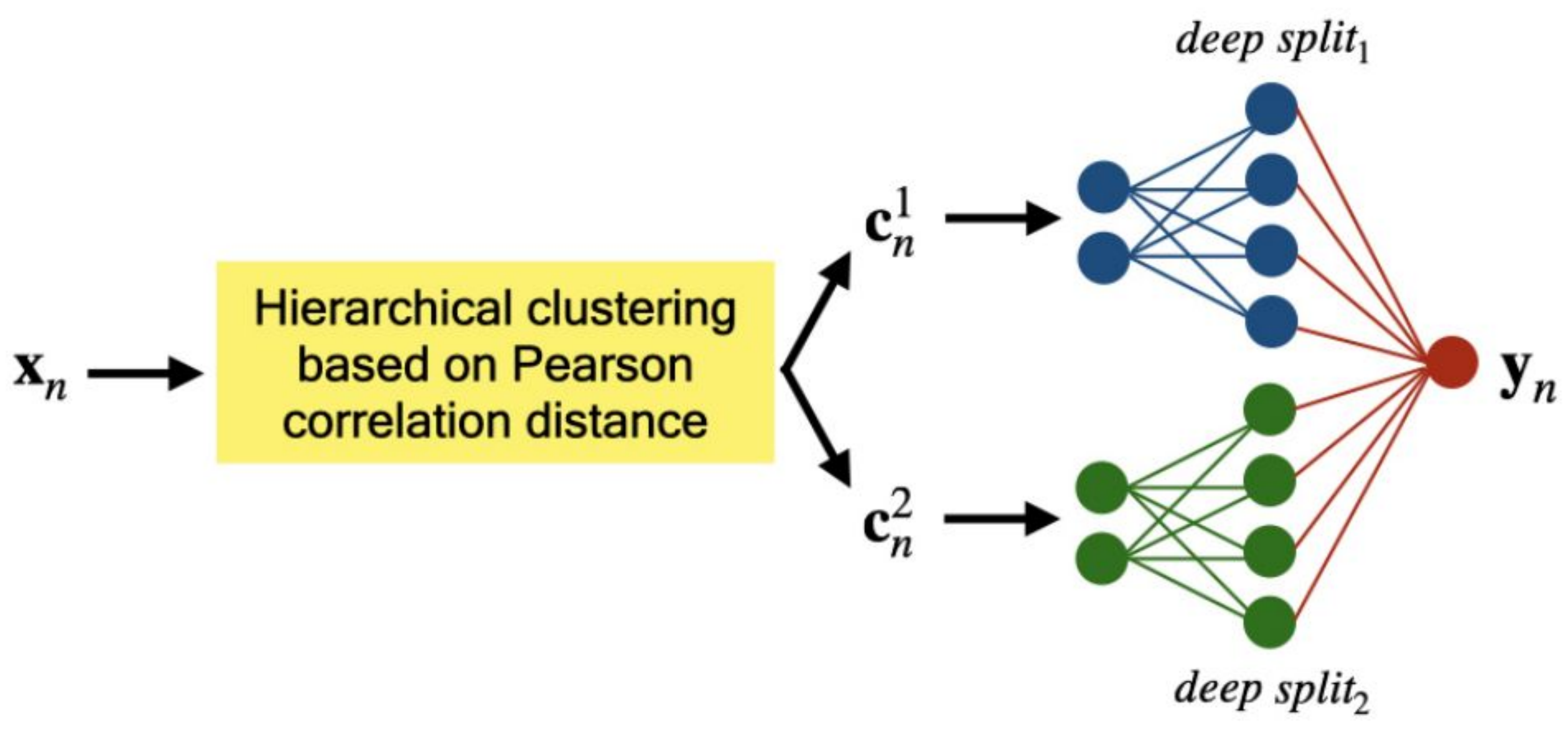
Rishab Khincha, Utkarsh Sarawgi, Wazeer Zulfikar, Pattie Maes

{rkhincha, utkarshs, wazeer, pattie} @ mit.edu

Summary

- The problem of missing data has been persistent for a long time and poses a major obstacle in machine learning and statistical data analysis.
- We propose an effective approach that clusters similar input features together using hierarchical clustering and then train proportionately split neural networks with a joint loss.
- We evaluate this approach on a series of benchmark datasets and show promising improvements even with simple imputation techniques.
- We attribute this to learning through clusters of similar features in our model architecture.

Process Diagram



Hierarchical clustering based on Pearson correlation distance

Procedure

- **Feature clustering**: The input feature space is split into *k* exhaustive clusters using hierarchical clustering based on Pearson correlation distance. Note that we are clustering features, which should not be confused with clustering data points. Splitting the input features in this way is effective since the cluster of similar features tend to work well together to substitute for the missingness, to provide better estimates.
- **Split NN:** The NN is then split with hidden units in each of the deep splits proportional to the number of features in the corresponding clusters, as shown in the process diagram. The split NN is then trained with all feature clusters using a joint loss (CCE for classification and MSE for regression), wherein the missing values are imputed for the mean value of that input feature.

Table 1: Dataset details

Dataset	Samples	Features	Missing	k
bands	539	19	5.38%	10
kidney disease	400	24	10.54%	9
hepatitis	155	19	5.67%	14
horse	368	22	23.80%	14
mammographics	961	5	3.37%	4
pima	768	8	12.24%	7
winconsin	699	9	0.25%	6
life expectancy	2938	21	43.7%	8

Evaluation and Results

- Our network consists of 50 hidden units with ReLU activations, trained to optimize for the categorical cross-entropy (CCE) loss.
- We use a 5-fold double cross-validation setup to report classification accuracies and train all the networks with a learning rate of 0.01 and a batch size of 100 for 1000 epochs.
- We evaluate on benchmark datasets (Table 1) from the literature and the ‘Life Expectancy (WHO)’ dataset and find that the results are competitive with other state-of-the-art methods even with simple imputation techniques like mean imputation (Table 2).
- While the performance of NNs usually drop, we observe that Split NNs are relatively robust to it as a consequence of statistically correlated features clustered together (Table 3).

Discussion

- Learning and inferring from data with incomplete features has been a pervasive problem in machine learning and statistical analysis.
- Creating clusters of statistically correlated input features show impressive performance even with using simple imputation techniques.
- We are very excited with the initial results and the future venues this work opens up.

Table 2: Classification accuracies on benchmark datasets

Dataset	karma	mice	mean	dropout	Smieja et al.	Vanilla NN	Split NN (ours)
bands	0.580	0.544	0.545	0.616	0.598	0.551 ± 0.058	**0.662 ± 0.051**
kidney disease	**0.995**	0.992	0.985	0.983	0.993	0.972 ± 0.030	0.963 ± 0.032
hepatitis	0.665	0.792	0.825	0.780	0.846	0.716 ± 0.069	**0.849 ± 0.075**
horse	0.826	0.820	0.793	0.823	**0.864**	0.794 ± 0.036	0.826 ± 0.020
mammographics	0.773	0.825	0.819	0.814	**0.831**	0.827 ± 0.026	**0.829 ± 0.016**
pima	0.768	0.769	0.760	0.754	0.747	0.762 ± 0.020	**0.777 ± 0.039**
winconsin	0.958	**0.970**	0.965	0.964	**0.970**	0.961 ± 0.015	0.964 ± 0.009

Table 3: RMSE scores on the Life Expectancy dataset (43.7% missing features)

Model	Val RMSE	Test RMSE
Vanilla NN	3.882	5.116
Split NN ($k = 8$)	2.945	4.246
Split NN ($k = 2$)	3.584	**4.006**

Paper

Code