Supplementary 2.

a). Performance of ICE is Robust in a Wide Range of Parameter Space

This supplementary file provides the analysis and discussion for the choice of the hyper-parameters. Figure 1 shows the grid search results of ICE using a wide range of hyper-parameters – N: number of neighbors per testing instances in prediction; w: the weight advantage of the base whole model in model-instance association; s: the weight advantage of the self-model in model-instance association. In Figure 1, the proportion of 'whole' model parameter α and β are both set to 1 for a stable performance.

Figure 1a shows that the number of nearest neighbors used in model selection stage has only slight impact on AUC gain on average across all 49 datasets. The recommended setting of N is 5 to 10 for a balanced running speed and accuracy. ICE works best when there are strong patterns in the dataset. If ICE does not have a significant gain over Random Forest (RF) on a center dataset, a larger N setting will make ICE more stable and closer to RF. ICE still has a large room of improvement on specific dataset by using more suitable fuzzy clustering algorithm, which is one of our future work.

Figure 1b and Figure 1c shows the robust performance of ICE with respect to parameter w and s. A general insight of w and s is to set s slightly larger than w, such as s = 0.5, w=0.4. The parameter α and β are quite simple to choose. Set both α and β to 1 will lead to a decent result for most of cases; try to set both α and β to 0 if there are strong clusters within the dataset, and the extreme localized classifiers may have an advantage over the basic to-go choice where α= β=1.

As stated in section 3.1 'Data and Experimental Setup', the hyper-parameters of ICE for the main evaluation (section3.3) are chosen intuitively following the simple rule of thumb, without exhaustive tuning.

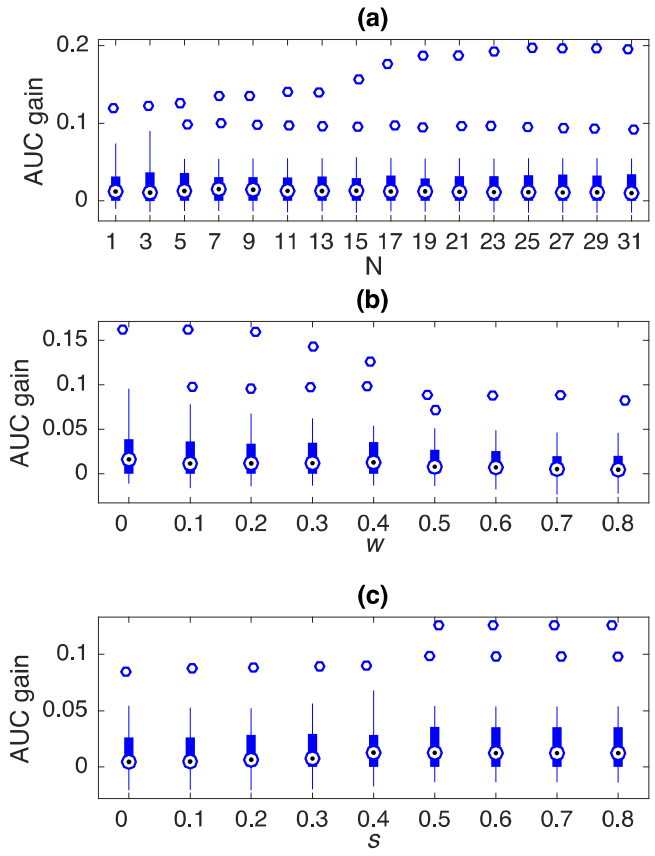

Figure 1
ICE performs in a stable manner across the wide range of parameter space. (a) AUC gain varies as a function of N, number of nearest neighbors for model selection. Here w = 0.4, s = 0.5. (b) AUC gain varies as a function of w. Here N = 5, s = 0.5. (c) AUC gain varies as a function of s. Here N =5, w=0.4. α=β=1 for all (a), (b) and (c)

b). Model Complexity

As in section 2.3, Q is the #instances; #cluster centers = ⌈10*log10(Q)⌉; each cluster center attaches to a few clusters with size z ∈ {2^i | 4<=i<=log2⌈3/4 * Q⌉}. For example, a dataset with 300 instances will have 25 centers; each center has 4 clusters, with size 2^4, 2^5 till 2^7. As in section 2.3, 'A classifier is then built using instances from each cluster', so there are 100 `partial' models built by ICE.

As shown in section 3.3, ICE with ~100 submodels outperforms RF with 10k trees. In fact, ICE only uses ~30 models per instance in testing for all datasets, where the complexity is much less comparing to the RF model with 10k trees. Different from RF, ICE builds up individualized models and selects the best models in prediction, rather than ensemble models built on random subset of instances and features. RF easily reaches its performance limits as the number of trees growing, while ICE has a much larger room of improvement as the number of submodels growing (Figure

4a). If we keep increasing the complexity of ICE by generating much more partial models (on many fuzzy clusters), the performance can be further improved.