A Novel Local Learning-Based Approach with Application to Breast Cancer Diagnosis

Songhua Xu, Ph.D. and Georgia Tourassi, Ph.D.

Biomedical Science and Engineering Center
Oak Ridge National Laboratory
Oak Ridge, TN, 37831

Purpose: The purpose of this study is to develop and evaluate a novel local learning-based approach for computer-assisted diagnosis of breast cancer.

Method: Our new local learning based algorithm using the linear logistic regression method as its base learner can be described as follows.

Step 1: Let G be the sample population that consists of n samples, i.e. $G = \{g_1, g_2, \ldots, g_n\}$. Each sample g_i carries 11 quantifiable features, represented as $f_j(g_i)$ $(j = 1, \ldots, 11)$. Given G, our algorithm first randomly selects a clustering scheme $\Phi(G)$ over G. In our implementation, we use the k-Nearest Neighbour (kNN) clustering algorithm to generate the random clustering scheme. This is done by randomly selecting the number of clusters, k, for the whole data set. Given k, we then randomly select k samples as the initial seeds to perform our kNN clustering process. In addition, we also stochastically search for a pairwise sample distance metric $\theta(g_i, g_s)$ through randomly assigning a series of weight parameters $\varpi = (\omega_1, \omega_2, \ldots, \omega_{11})$ such that $\theta(g_i, g_s) = \sum_{j=1}^{11} \omega_j (f_j(g_i) - f_j(g_s))$.

Step 2: Under the clustering scheme $\Phi(G)$, we partition the whole sample population into several sub-populations $G_1, G_2, \ldots, G_k$ such that $G = \bigcup_{l=0}^{k} G_l$ and $G_s \bigcap G_t = \varnothing$ $(s \neq t)$. For each such sub-population G_i, we then train a base learner L_i, which in our current implementation is a linear logistic regression model. All the trained base learners coupled with the clustering scheme $\Phi(G)$ then form our local learning model for the input entire population G, denoted as $M_{\Phi(G)}$.

Step 3: We iterate between steps 1 and 2 above. For each trained model instance $M_{\Phi(G)}$ from step 2, we test its performance according to the validation part of the input data set for model selection purpose. Note that the testing part of the input data set will not be utilized throughout the whole training process. To measure the performance of a trained model instance, we use the prediction area under curve value (AUC) as the performance metric. Our algorithm also keeps track of the performance of all the model instances derived so far at any moment of our algorithm running time. During our stochastic clustering schema searching process, we also keep track of the collective performance of a certain clustering sampling configuration in terms of the number of sub-populations k it divides the whole population into and the distance weight parameters ϖ. We measure the collective performance of a clustering sampling configuration using the best prediction performance of our local learning model $M_{\Phi(G)}$ derived using one of its yielded clustering scheme $\Phi(G)$. The higher the collective performance value is, the more likely a similarly clustering configuration will be sampled in the subsequent iterations. In measuring the similarity

between two clustering configuration, we use the following metric: $Dist(conf_1, conf_2) = 10^5 \,|\, k_{conf_1} - k_{conf_2} \,| + \| \omega_{conf_1} - \omega_{conf2} \|$, where $\|.\|$ denotes the Euclidean norm. Overall, our algorithm will perform its stochastic searching process until the total allowed computing time is used up by our random walk process in identifying the most suitable population subdivision scheme and their corresponding individual base learners.

€

The proposed local learning-based approach was applied for the prediction of breast cancer given 11 mammographic and clinical findings reported by physicians using the BI-RADS lexicon. Our database consisted of 850 patients with biopsy confirmed diagnosis (290 malignant and 560 benign). We also compared the performance of our method with a collection of publicly available state-of-the-art machine learning methods.

Results: Predictive performance for all classifiers was evaluated using 10-fold cross validation and Receiver Operating Characteristics (ROC) analysis. Figure 1 reports the performance of 54 machine learning methods implemented in the machine learning toolkit Weka (version 3.0). These methods include: 1) Bayesian Logistic Regression, 2) Naïve Bayes, 3) Naïve Bayes Simple, 4) Naïve Bayes Updateable, 5) Logistic, 6) Multilayer Perceptron, 7) RBF Network, 8) Simple Logistic Regression, 9) Nested Dichotomies, 10) Filtered Classifier, 11) Grading, 12) Decision Stump, 13) LMT, 14) Simple Cart, 15) Ada Boost, 16) Attribute Selected Classifier, 17) Bagging, 18) Classification Via Clustering, 19) Classification Via Regression, 20) CV Parameter Selection, 21) Dagging, 22) J48 Tree, 23) Logit Boost, 24) Multi Boost AB, 25) Multi Class Classifier, 26) FT Tree, 27) NB Tree, 28) REP Tree, 29) Bayes Net, 30) SVM (Poly Kernel), 31) SPegasos, 32) Voted Perceptron, 33) IB1, 34) Linear NN Search, 35) KStar, 36) LWL (Decision Stump), 37) Multi Scheme, 38) Hyper Pipes, 39) VFI, 40) J48 graft, 41) Random Forest, 42) Conjunctive Rule, 43) Decision Table, 44) DTNB, 45) JRip, 46) NNge, 47) One R, 48) PART, 49) Ridor, 50) Zero R, 51)AD Tree, 52) BF Tree, 53) LAD Tree, 54) Random Tree.

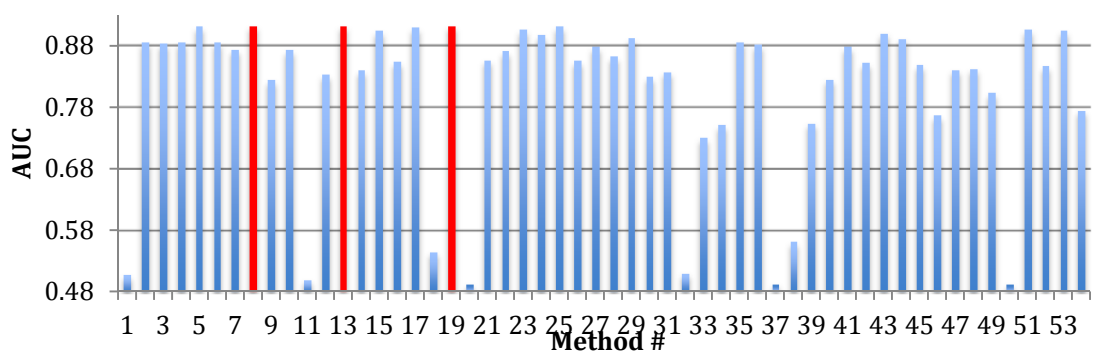


Figure 1. Performance comparison of 54 machine learning methods for our breast cancer prediction problem. See text.

To the best of our knowledge, Weka's implementation of cross-validation is based on randomly dividing the whole sample population in a way that is fixed for all methods and all runs. Therefore, performance measurement numbers obtained for different methods can be directly compared. The best prediction performance observed is 0.912 (as determined by Weka's ROC implementation), which is attained by three methods independently: Simple

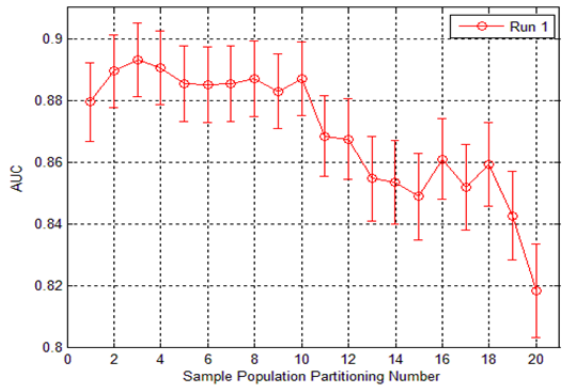
Logistic Regression, LMT, and Classification Via Regression—all highlighted in red in the figure. This finding is consistent with prior studies confirming the highly linear nature of the problem in that simple linear regression is capable of achieving top performance among all popular machine learning methods. Our study results further confirm that using sophisticated machine learning approaches such as multi-layer perceptron, Adaboost, and multi class classifier do not provide any further improvement. We believe that the more sophisticated decision boundaries supported by these advanced learning methods cannot effectively help improve the learning performance, but only subject the methods to higher overfitting risk.

Figure 2 shows the corresponding performance for the proposed approach in terms of its AUC and the partial AUC value ($pAUC_{TPF0.9,\ 1.0}$) for the case when our local learning method partitions the whole sample population into different numbers of sub-populations sp#=1, …, 20. Note that sp#=1 corresponds to a degenerated case where no local learning scheme is used and the entire sample population is learned as a whole. This setting provides the baseline method (i.e., simple linear regression). We used the Matlab function call of linear logistic regression to realize our base learner and the ROCKIT software to compute both AUC and $pAUC_{TPF0.9,\ 1.0}$ values. The figure shows the results of two different ten-fold sample division plans, demonstrating the stability of our study conclusion independent from any random ten-fold sample division plan. As the figure indicates, our local learning method outperforms the baseline linear logistic regression method with statistical significance at the 95% confidence level for both runs.

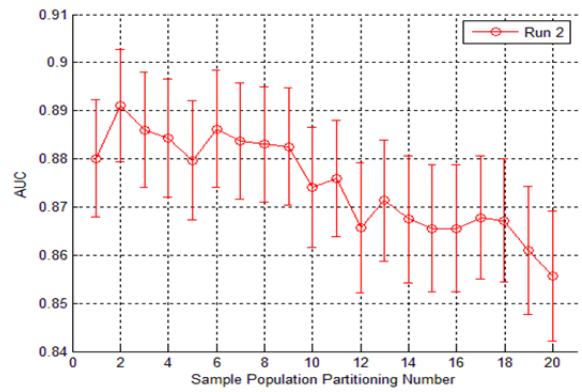
Please note that the AUC differences of the simple logistic regression method between Figures 1 and 2 could be easily attributed to differences in the implementation of the 10-fold cross validation scheme and the software used to estimate the ROC area. The results shown in Figure 1 are based on the Weka software, which does not output its ten fold sample data split for us to employ in our own experiments. The results shown in Figure 2 are based on in-house software and the ROCKIT software for estimating AUCs and partial pAUCs due to the more accurate estimation algorithm it implements. Due to these differences, the numbers reported in Figures 1 and 2 can not be directly compared. However, the qualitative conclusions remain: simple linear logistic regression achieves the best performance among a wide range of sophisticated machine learning methods implemented in Weka, yet our local learning approach achieves a noticeable and statistically meaningful performance improvement, which is numerically validated through a set of comparison experiments.

New or breakthrough aspect of work: Our experimental results suggest that it is worth exploring local learning techniques even when tackling problems of highly linear structure. This conclusion complements the existing results in the machine learning field that local learning may work effectively in capturing complicated, non-linear relationships exhibited by real-world datasets.

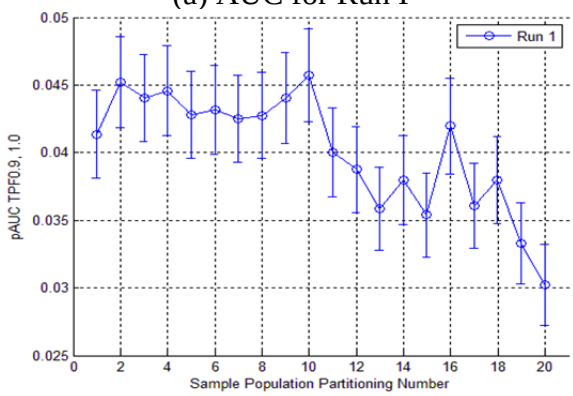
Conclusion: We introduced a novel local learning-based classifier and compared it with an extensive list of other classifiers for the problem of breast cancer diagnosis. Our experiments show that the algorithm superior prediction performance outperforming a wide range of other well established machine learning techniques. Our conclusion complements the existing understanding in the machine learning field that local learning may capture complicated, non-linear relationships exhibited by real-world datasets.



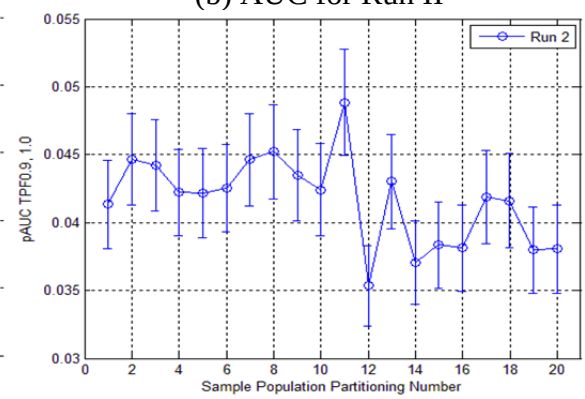
(a) AUC for Run I



(b) AUC for Run II



(c) $pAUC_{TPF0.9,\ 1.0}$ for Run I



(d) $pAUC_{TPF0.9,\ 1.0}$ for Run II

Runs	Run I	Run II
P_{base}	0.8795±0.0126	0.8801±0.0122
P_{our}	**0.8911±0.0119**	**0.8911±0.0116**
$P_{our\text{-}base}$	0.0211	0.0104

(e) Comparison between the performance of our global base learner (P_{base}), overall performance of our local learning method (P_{our}), and the P-value of our method's performance against that of the global base learner ($P_{our\text{-}base}$).

Figure 2. Performance analysis and comparison of our local learning method with respect to its base learner.

Acknowledgement This work has never been presented or submitted for publication elsewhere. This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.