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

Songhua Xu[1], Ph.D. and Georgia Tourassi[2], Ph.D.

Biomedical Science and Engineering Center,
Oak Ridge National Laboratory,
Oak Ridge, Tennessee 37831, USA

ABSTRACT

In this paper, we introduce a new local learning based approach and apply it for the well-studied problem of breast cancer diagnosis using BIRADS-based mammographic features. To learn from our clinical dataset the latent relationship between these features and the breast biopsy result, our method first dynamically partitions the whole sample population into multiple sub-population groups through stochastically searching the sample population clustering space. Each encountered clustering scheme in our online searching process is then used to create a certain sample population partition plan. For every resultant sub-population group identified according to a partition plan, our method then trains a dedicated local learner to capture the underlying data relationship. In our study, we adopt the linear logistic regression model as our local learning method's base learner. Such a choice is made both due to the well-understood linear nature of the problem, which is compellingly revealed by a rich body of prior studies, and the computational efficiency of linear logistic regression--the latter feature allows our local learning method to more effectively perform its search in the sample population clustering space. Using a database of 850 biopsy-proven cases, we compared the performance of our method with a large collection of publicly available state-of-the-art machine learning methods and successfully demonstrated its performance advantage with statistical significance.

Keywords: mammography, classification, local-learning, breast cancer diagnosis prediction.

[1] Email: xus1@ornl.gov.

[2] Email: tourassig@ornl.gov.

Notice: This manuscript has been authored by UT-Battelle, LLC, under Contract No. 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.

1. INTRODUCTION & RELATED WORK

The purpose of this study is to develop and evaluate a novel local learning-based approach for computer-assisted diagnosis of breast cancer. The idea of local learning dates back more than a decade ago [2,3,12], initially proposed by theoretical machine learning researchers. Due to the fast growth of computing power available, local learning has recently become truly affordable for dealing with real-world problems with large data sets. Outside the medical world, the idea of local learning has been successfully applied for modeling and mining industry data sets. For example, Kadlec and Gabrys introduced a local learning based approach for soft sensor data modeling [1]. Yoon and Cho [4] used hybrid global and local learners to emulate a mixed panel of experts for learning data labels. Hartono and Hashimoto adopted the local learning idea to produce an ensemble of neural networks [5] and also an ensemble of linear perceptrons [6] for learning data selection mechanisms. Dong et al. [7] introduced a local learning framework for recognition of lowercase handwritten characters. Qin and Forbes [8] utilized the local learning idea for executing dynamic regional harmony search. Nadeem and Fahringer [9] applied local learning for predicting the execution time of grid workflow applications. Sun and Wu [10] used local learning to compile a set of quality features and proved the effectiveness of their approach through experiments both on synthetic and real-world data sets. Sun et al. [11] used the local learning idea for feature selection in analyzing high-dimensional data. Li et al. [13] introduced a specialized local learning technique for predicting queue wait time. Despite the plethora of successes witnessed regarding the idea of local learning in physical sciences and engineering, the power of local learning has yet to be realized in the medical imaging world. In this paper, we introduce a new local learning based approach and apply it for the well-studied problem of characterizing the malignancy status of breast masses using BIRADS-based mammographic features [14].

2. OUR METHOD

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

<u>Step 1:</u> Let G be the sample population that consists of n samples, i.e. $G=\{g_1,g_2,\cdots,g_n\}$. Each sample g_i carries 11 quantifiable features, represented as $f_j(g_i)$ $(j=1,\cdots,11)$. 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 $\varpi=(\omega_1,\omega_2,\cdots\omega_{11})$ such that $\theta(g_i,g_s)=\sum_{j=1}^{11}\omega_j(f_j(g_i)-f_j(g_s))^2$.

Step 2: Under the clustering scheme $\Phi(G)$, we partition the whole sample population into several sub-populations $G_1, G_2, \cdots, G_k$ 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 trained base learners coupled with the clustering scheme $\Phi(G)$ then form our local learning model for the entire input 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 the model selection purpose. Note that the testing part of the input data set is not utilized throughout the whole training process. To measure the performance of a trained model instance, we apply Receiver Operating Characteristics (ROC) analysis and use the prediction ROC area under curve value (AUC) as the performance metric [15]. Our algorithm also keeps track of the performance of all model instances derived 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 and the weight parameters ϖ. We measure the collective performance of a clustering sampling configuration using the best prediction AUC performance of our local learning model $M_{\Phi(G)}$ derived using one of its yielded clustering schemes $\Phi(G)$. Note that a clustering configuration $\Phi(G)$ implies a specific combination of cluster number *k*, weight vector ϖ, and initialization seed(s). The higher the collective performance value is, the more likely a similar clustering configuration will be sampled in the subsequent iterations. In measuring the similarity between two clustering configurations, 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 performs 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.

3. EXPERIMENT RESULTS

The proposed local learning-based approach was applied for predicting the malignancy status of breast masses based on 11 features: 5 mammographic (mass margin, mass shape, mass density, mass size, associated architectural distortion) reported by radiologists using the BI-RADS lexicon, 5 clinical findings (patient age, family history of breast cancer, history of hormonal therapy, history of BRCA, menopausal status), as well as the radiologists' BI-RADS assessment of malignancy. Our database consisted of 850 biopsy-proven breast masses (290 malignant and 560 benign). Based on their BI-RADS assessment, the radiologists' AUC diagnostic performance was 0.8573±0.0144.

We compared the performance of our method with a collection of publicly available state-of-the-art machine learning methods. Table 1 shows the detailed list of the 54 machine learning methods we tested with their corresponding AUCs estimated by 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, and 54) Random Tree. Figure 1 illustrates the AUC performance of all 54 machine learning methods. The reported performance was based on 10-fold cross validation performed using Weka's own implementation.

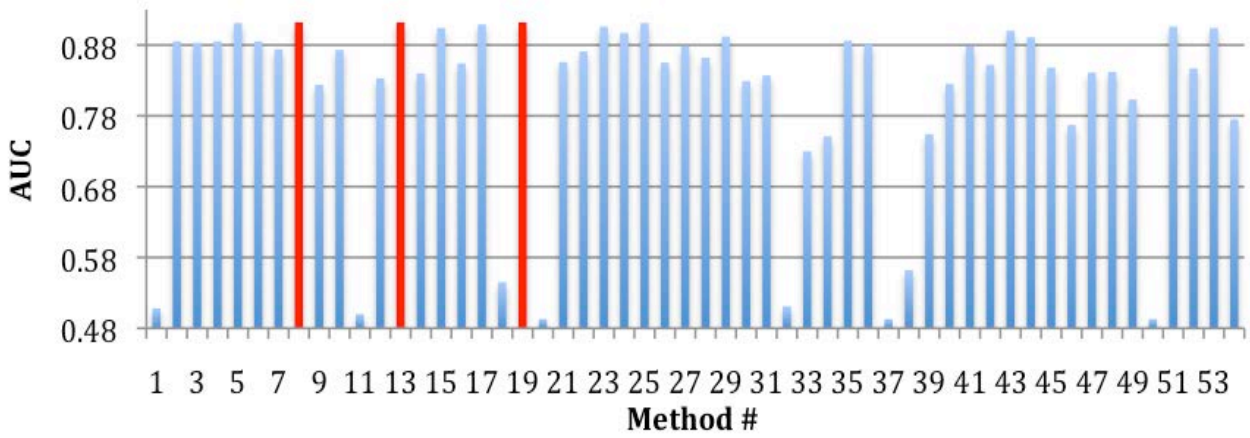


Figure 1. AUC performance comparison of 54 machine learning methods for our breast cancer diagnosis problem.

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 AUC implementation), which is attained by three different methods: Simple Logistic Regression, LMT, and Classification Via Regression—all highlighted in red in Figure 1. 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 produced by these advanced learning methods cannot effectively improve the learning performance, but only subject the methods to higher overfitting risk.

Table 1. AUC performance comparison of 54 machine learning methods for breast cancer diagnosis.

Method	AUC	Method	AUC	Method	AUC	Method	AUC
1. Bayesian Logistic Regression	0.508	15. Ada Boost	0.904	29. Bayes Net	0.892	42. Conjunc-tive Rule	0.852
2. Naïve Bayes	0.885	16. Attribute Selected Classifier	0.854	30. SVM (Poly Kernel)	0.829	43. Decision Table	0.900
3. Naïve Bayes Simple	0.883	17. Bagging	0.909	31. SPegasos	0.837	44. DTNB	0.891
4. Naïve Bayes Updateable	0.885	18. Classifica-tion Via Clustering	0.545	32. Voted Perceptron	0.511	45. JRip	0.848
5. Logistic	0.911	**19. Classifica-tion Via Regression**	**0.912**	33. IB1	0.730	46. NNge	0.767
6. Multilayer Perceptron	0.885	20. CV Parameter Selection	0.493	34. Linear NN Search	0.751	47. One R	0.841
7. RBF Network	0.874	21. Dagging	0.856	35. KStar	0.886	48. PART	0.842
8. Simple Logistic Regression	**0.912**	22. J48 Tree	0.871	36. LWL (Decision Stump)	0.881	49. Ridor	0.803
9. Nested Dichotomies	0.824	23. Logit Boost	0.906	37. Multi Scheme	0.493	50. Zero R	0.493
10. Filtered Classifier	0.873	24. Multi Boost AB	0.897	38. Hyper Pipes	0.562	51. AD Tree	0.906
11. Grading	0.500	25. Multi Class Classifier	0.911	39. VFI	0.754	52. BF Tree	0.847
12. Decision Stump	0.833	26. FT Tree	0.855	40. J48 graft	0.825	53. LAD Tree	0.904
13. LMT	**0.912**	27. NB Tree	0.878	41. Random Forest	0.878	54. Random Tree	0.775
14. Simple Cart	0.84	28. REP Tree	0.862	**Best prediction performance: 0.912,** attained by Classification Via Regression, LMT, and Simple Logistic Regression.			

Since linear logistic regression was one of the best performing and simplest Weka machine learning methods, we compared our approach to it using the same cross-validation plan. Table 2 shows the corresponding performance for the proposed approach in terms of its AUC and the partial AUC value ($_{0.90}$AUC) for the case when our local learning method partitions the whole sample population into different numbers of sub-populations for k=1, …, 20.

Table 2. AUC performance analysis and comparison of our local learning method with respect to its base learner.

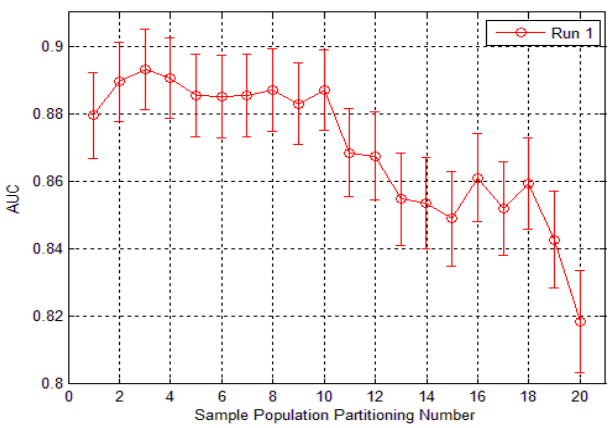
	Run 1		**Run 2**	
k	**AUC**	**$_{0.90}$AUC**	**AUC**	**$_{0.90}$AUC**
1	0.8795 ± 0.0126	0.0414±0.0032	0.8801±0.0122	0.0414±0.0033
2	0.8895±0.0118	0.0452±0.0034	**0.8911±0.0116**	0.0447±0.0034
3	**0.8932±0.0119**	0.0440±0.0032	0.8860±0.0120	0.0442±0.0034
4	0.8905±0.0119	0.0446±0.0033	0.8843±0.0122	0.0422±0.0032
5	0.8855±0.0122	0.0428±0.0032	0.8797±0.0124	0.0422±0.0033
6	0.8851±0.0121	0.0432±0.0033	0.8862±0.0122	0.0426±0.0032
7	0.8853±0.0122	0.0425±0.0032	0.8837±0.0121	0.0446±0.0034
8	0.8870±0.0121	0.0428±0.0032	0.8830±0.0120	0.0452±0.0035
9	0.8829±0.0121	0.0440±0.0034	0.8825±0.0122	0.0435±0.0033
10	0.8871±0.0119	0.0457±0.0035	0.8741±0.0125	0.0424±0.0034
11	0.8685±0.0129	0.0400±0.0033	0.8759±0.0120	0.0489±0.0039
12	0.8675±0.0131	0.0388±0.0032	0.8656±0.0135	0.0353±0.0029
13	0.8547±0.0137	0.0359±0.0031	0.8714±0.0126	0.0430±0.0035
14	0.8535±0.0136	0.0380±0.0033	0.8675±0.0132	0.0371±0.0030
15	0.8489±0.0139	0.0354±0.0031	0.8655±0.0132	0.0384±0.0032
16	0.8611±0.0130	0.0420±0.0036	0.8655±0.0132	0.0381±0.0032
17	0.8519±0.0138	0.03609±0.00315	0.8678±0.0128	0.0419±0.0034
18	0.8593±0.0134	0.03797±0.00323	0.8672±0.0128	0.0416±0.0034
19	0.8427±0.0143	0.03331±0.00301	0.8610±0.0133	0.0380±0.0032
20	0.8185±0.0151	0.03024±0.00298	0.8556±0.0135	0.0381±0.0033
Overall	**0.8932±0.0119**		**0.8911±0.0116**	
P-value	0.0211		0.0104	

Note that k=1 corresponds to a degenerate case where no local learning scheme is used and the entire sample population is learned as a whole. This setting provides the baseline method where logistic regression alone is used. The table includes detailed results for two separate runs of the same experiment (Runs 1 and 2). These runs represent two different ten-fold sample division plans but the conditions for both runs are kept otherwise the same. The second to last row of the table, titled "overall," reports the overall best performance of our local learning method across all k's in terms of AUC. The last row of the table shows the two-tailed p-value for the statistical comparison between the overall performance of our local learning method to that of the baseline linear logistic regression method. As the table shows, our approach performs statistically significantly better than the baseline linear logistic regression method. Furthermore, a small number of sub-populations (k=3 for Run 1 and k=2 for Run 2) appears to be the optimal for the specific problem.

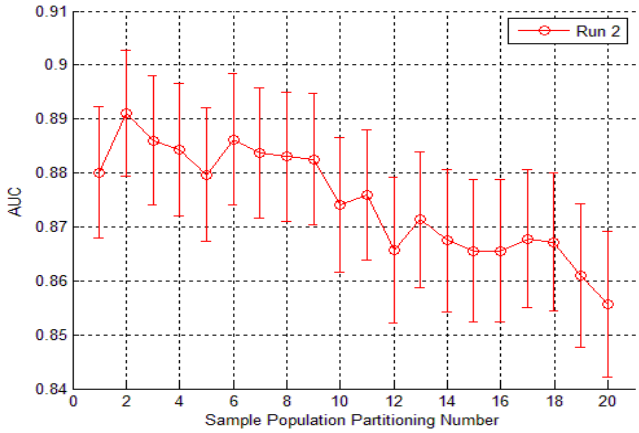
Figure 2 illustrates these performance testing results. We used the Matlab function call of linear logistic regression to realize our base learner and the ROCKIT software to compute both AUC and $_{0.90}$AUC values. The figure shows the results of two different runs, demonstrating the stability of our 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. 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 AUC 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 these differences, the numbers reported in Figures 1 and 2 cannot be directly compared. However, the qualitative conclusions remain the same: 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 significant performance improvement

4. CONCLUSION

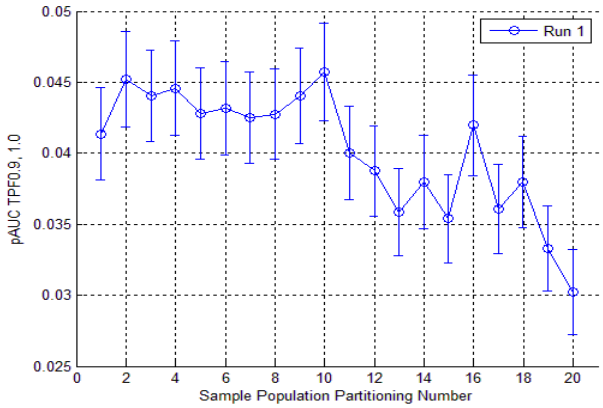
We introduced a novel local learning-based classifier and compared it with an extensive list of other classifiers for the problem of breast cancer diagnosis. Our experiments showed that our classification algorithm had superior prediction performance, outperforming a wide range of other well established machine learning techniques for the problem of breast cancer diagnosis. Despite the well-known linear nature of the problem, our local learning approach achieved a performance improvement, which was quantitatively validated through a set of two comparison experiments. Besides the superior machine learning method readily offered by our novel local learning approach, our experimental results also suggest that it is worth exploring local learning techniques even when tackling problems of highly linear structure. This conclusion complements the existing understanding in the machine learning community that local learning may capture complicated, non-linear relationships exhibited in real-world datasets.



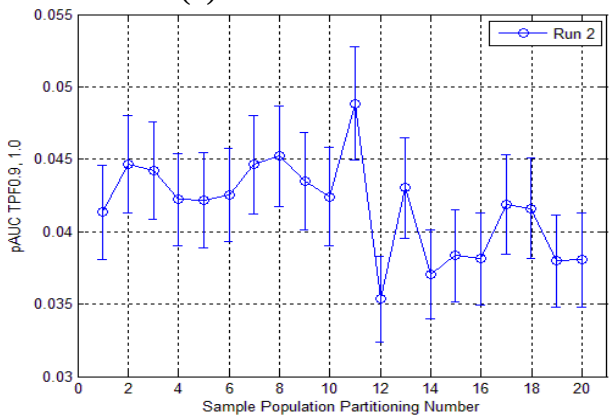
(a) AUC for Run 1



(b) AUC for Run 2



(c) pAUC for Run 1



(d) pAUC for Run 2

Runs	Run 1	Run 2
AUC_{base}	0.8795±0.0126	0.8801±0.0122
AUC_{our}	**0.8932±0.0119**	**0.8911±0.0116**
2-tailed P-value	0.0211	0.0104

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

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

ACKNOWLEDGEMENT

Songhua Xu performed this research as a Eugene P. Wigner Fellow and staff member at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

REFERENCES

[1] Petr Kadlec and Bogdan Gabrys. 2008. Soft sensor based on adaptive local learning. In Proceedings of the 15th International Conference on Advances in Neuro-Information Processing - Volume Part I (ICONIP'08), Vol. I. Springer-Verlag, Berlin, Heidelberg, 1172-1179.

[2] Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. 1997. Locally Weighted Learning. Artificial Intelligence Review 11(1-5): 11-73.

[3] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. Neural Computing 3(1): 79-87.

[4] Jong-Won Yoon and Sung-Bae Cho. 2011. Global/local hybrid learning of mixture-of-experts from labeled and unlabeled data. In Proceedings of the 6th International Conference on Hybrid Artificial Intelligent Systems – Volume I, Springer-Verlag, Berlin, Heidelberg, 452-459.

[5] Pitoyo Hartono and Shuji Hashimoto. 2001. Learning data selection mechanism through neural networks ensemble. In Proceedings of the Second International Workshop on Multiple Classifier Systems (MCS '01). Springer-Verlag, London, UK, 188-197.

[6] Pitoyo Hartono and Shuji Hashimoto. 2005. Learning with ensemble of linear perceptrons. In Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications - Volume II. Springer-Verlag, Berlin, Heidelberg, 115-120.

[7] Jian-Xiong Dong, Adam Krzyzak, and Ching Y. Suen. 2001. Local learning framework for recognition of lowercase handwritten characters. In Proceedings of the Second International Workshop on Machine Learning and Data Mining in Pattern Recognition (MLDM '01). Springer-Verlag, London, UK, 226-238.

[8] A. K. Qin and Florence Forbes. 2011. Dynamic regional harmony search with opposition and local learning. In Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation (GECCO '11). ACM, New York, NY, USA, 53-54.

[9] Farrukh Nadeem and Thomas Fahringer. 2009. Predicting the execution time of grid workflow applications through local learning. In Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis (SC '09). ACM, New York, NY, USA, Article 33, 12 pages.

[10] Yijun Sun and Dapeng Wu. 2009. Feature extraction through local learning. Statistical Analysis and Data Mining, 2(1): 34-47.

[11] Yijun Sun, Sinisa Todorovic, and Steve Goodison. 2010. Local-learning-based feature selection for high-dimensional data analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9), 1610-1626.

[12] Santosh Ananthraman. 1993. Robot system identification and control using a rapid local-learning artificial neural network paradigm. Ph.D. Dissertation. Duke University, Durham, NC, USA.

[13] Hui Li, Juan Chen, Ying Tao, David Gro, and Lex Wolters. 2006. Improving a local learning technique for queuewait time predictions. In Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID '06). IEEE Computer Society, Washington, DC, USA, 335-342.

[14] J. Y. Lo, A. O. Bilska-Wolak, M. K. Markey, G. D. Tourassi, J. A. Baker, and C. E. Floyd Jr. 2006. Computer-aided diagnosis in breast imaging: where do we go after detection? In Recent Advances In Breast Imaging, Mammography, and Computer-Aided Diagnosis of Breast Cancer (eds. Suri and Rangayyan), 871-900.

[15] G. D. Tourassi. 2010. Receiver operating characteristics analysis: Basic concepts and practical application. Handbook of Medical Image Perception and Techniques, Cambridge University Press, Cambridge, UK.