GA-EoC: A Genetic Algorithm-based Ensemble Method for Enhancing the Dataset Classification Accuracy

Mohammad Nazmul Haque[1], Pablo Moscato[1,2], Regina Berratta[1], Manuel Ujaldon Martinez[1]

1 Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine and Hunter Medical Research Institute, School of Electrical Engineering and Computer Science, The University of Newcastle, University Drive, Callaghan NSW 2308, Australia.

2 Australian Research Council Centre of Excellence in Bioinformatics, Callaghan, New South Wales, Australia

Preliminaries:

Classification: A *Supervised Learning* method that learns a function from *training data* and *predict* discrete output labels for unknown dataset.

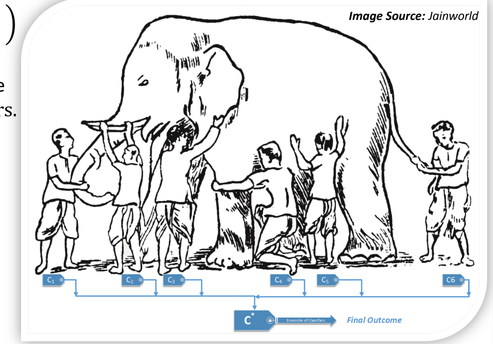
$$\mathbb{C}: \mathfrak{R}^n \underset{\mathrm{T}}{\rightarrow} \Omega$$

Here, $\mathbb{C}$=Classifier, $\mathfrak{R}^n$=n-dimensional feature space, T=Training datset and Ω=set of class labels.

Ensemble Method: A set of learning machines to learn partial solutions and integrate those to construct a final outcome to the original problem.

$$\mathbb{E}(\mathbb{C}^*)=\sum_{i=1}^{k} \omega_i \left(\mathbb{C}: \mathfrak{R}^n \underset{\mathrm{T}}{\rightarrow} \Omega\right)$$

Here, $\mathbb{E}$=Ensemble and $\omega_i \in \mathbb{R}$ is the measure of the goodness of classifiers.



Performance Measures:

		Actual	
		Pos	*Neg*
Predicted	*Pos*	TP	FP
	Neg	FN	TN

The Matthews Correlation Coefficient (MCC) often provide a much more balanced evaluation of the classifier accuracy [1]

$$\mathbf{MCC} = \frac{(\mathbf{TP}\times\mathbf{TN})-(\mathbf{FP}\times\mathbf{FN})}{\sqrt{(\mathbf{TP}+\mathbf{FP})(\mathbf{TP}+\mathbf{FN})(\mathbf{TN}+\mathbf{FP})(\mathbf{TN}+\mathbf{FN})}}$$

Rationale for the Ensemble of Classifiers (EoC) Method:

By using the Bernoulli trial formula [2], we get the probability (P) of observing r successes (x=n-r fails) in n trials as :

$$\boldsymbol{P(rS, xF)} = \binom{\boldsymbol{n}}{\boldsymbol{r}} \boldsymbol{p^r(1-p)^{(n-r)}}$$

If we have k single classifiers to form the ensemble where at more than half of them predict correctly, then the probability of success for the ensemble is:

$$\boldsymbol{P(\mathbb{E})} = \sum_{\boldsymbol{i}}^{\boldsymbol{k}} \binom{\boldsymbol{k}}{\boldsymbol{i}} \boldsymbol{p^i(1-p)^{(k-i)}}$$

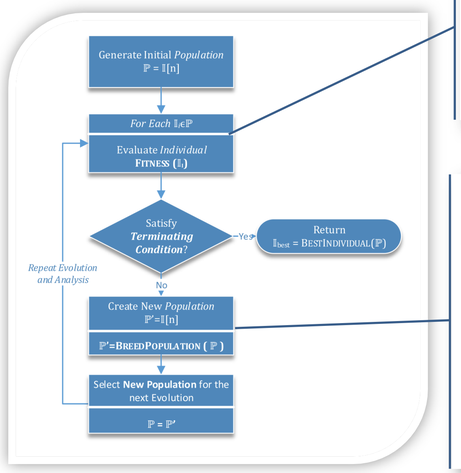
Let, $k = 20$ classifiers, individual prediction accuracy $p = 0.6$ and at least $i = 11$ of them predict correctly $(i > k/2)$. Then the ensemble accuracy:

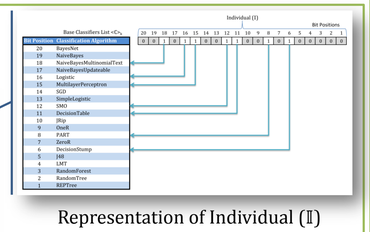
$$\boldsymbol{P(\mathbb{E})} = \sum_{\mathbf{11}}^{\mathbf{20}} \binom{\mathbf{20}}{\mathbf{11}} \mathbf{0.6^{11}(1-0.6)^{(20-11)}} = \mathbf{0.94347} \approx \mathbf{94\%}$$

The proposed Genetic Algorithm-based EoC:

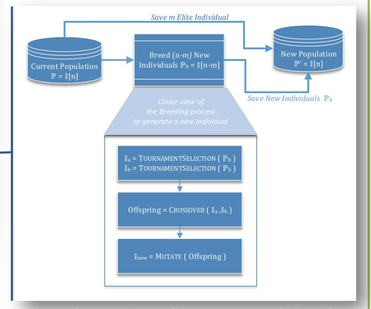
Based on the rationale of ensemble, we have chosen 20 single classifiers from the WEKA data mining software suite [3]. Which gives 2.43×10^{18} possible combinations for forming the ensemble.

The Proposed GA-EoC:





Representation of Individual (𝕀)



Population Breeding Process (ℙ'= ℙ)

Parameters of the Genetic Algorithm:

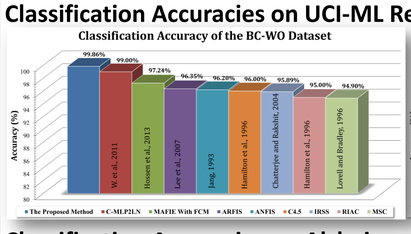
- **Objective Function :** $obj = arg \quad max_{i=1..j} \quad fit(\mathbb{I}_i \in \mathbb{P})$
- **Population Size:** $|\mathbb{P}|$ =100
- **Mutation Rate :** $\mathbb{R}_{mt} = 0.01$
- **Uniform Crossover :** $\mathbb{R}_{xo} = 0.60$
- **Terminating Conditions:**
  1. #generation reaches 10,000.
  2. fitness remains stationary for consecutive 50 generations.
  3. obj value reaches at 1.0

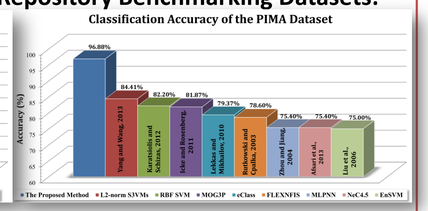
Details of Datasets:

Dataset	#Samples	#Features	Source
BC-WO	699	9	UCI-ML Repository
PIMA	768	8	UCI-ML Repository
Test Set AD	92	120	Ray et al.,2007 [4]
Test Set MCI	47	120	Ray et al.,2007 [4]
Ray-AD-Trn-18	83	18	Ray et al.,2007 [4]
RavettiMoscato-AD-Trn-5	83	5	Ravetti and Moscato, 2008 [5]

Experimental Results:

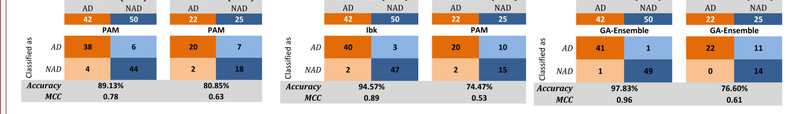
Classification Accuracies on UCI-ML Repository Benchmarking Datasets:





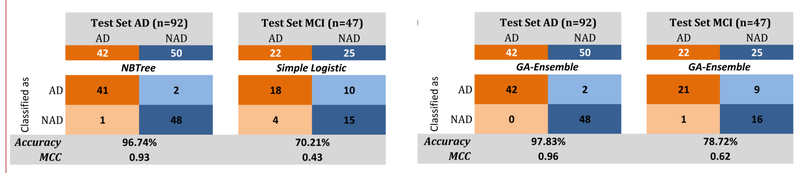
Classification Accuracies on Alzheimer's Disease Datasets:

1. Using *Ray-AD-Trn-18* (the 18-Protein Biomarker) as training dataset:



a) Ray et al., 2007 b) Ravetti and Moscato, 2008 c) The Proposed GA-EoC

2. Using *RavettiMoscato-AD-Trn-5* (5-Protein Biomarker) as training dataset:



a) Ravetti and Moscato, 2008 b) The Proposed GA-EoC

Conclusion:

The experimental results of the proposed method are promising. However, the current implementation works on binary-class datasets. We have to improve it by bringing the multi-class classification capability.

Key References:

[1] Dutt, R. and Madan, A. (2012). Predicting biological activity: Computational approach using novel distance based molecular descriptors. Computers in Biology and Medicine, 42(10):1026-1041

[2] Kobayashi, H., Mark, B. L., and Turin, W. (2011). *Probability, Random Processes, and Statistical Analysis*, chapter 2, pages 26-29. Cambridge University Press.

[3] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). *The weka data mining software: an update*. SIGKDD Explor. Newsl., 11(1):10-18

[4] Ray, S., Britschgi, M., Herbert, C., Takeda-Uchimura, Y., Boxer, A., Blennow, K., Friedman, L. F., Tibshirani, R., et al.. (2007). Classication and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins. Nature medicine, 3(11):1359-62.

[5] Ravetti, M. G. and Moscato, P. (2008). Identication of a 5-protein biomarker molecular signature for predicting Alzheimer's disease. PloS one, 3(9):e3111.