EECE5645 Proposal: Genetic Algorithms for Feature Selection

Potter, Mıchael   Yildiz, Ayberk Yarkın   Gordon, Cameron

Marer Prabhu, Nishanth

Motivation: Many technology domains such as hyperspectral imagery, microarrays, and the internet have created datasets with hundreds to hundreds of thousands of features [4]. Machine Learning tasks with high dimensionality pose challenges such as memory consumption, compute time, generalizability, and interpretability [4]. Thus, selecting the most informative subset of features is desirable. However, the cardinality of all combinatorial feature subsets to search is 2^d, which quickly becomes larger than the estimated number of atoms in the entire known universe for $d > 270$.

Proposal: We aim to parallelize the Genetic Algorithm (GA) for feature selection [2]. We can distribute the 'chromosomes' at each population iteration to enable concurrent training of Machine Learning (ML) models on diverse feature subsets, rather than sequentially training on a single device. Furthermore, the computation speed will enable an increase in the population size increasing the odds of finding the optimal feature subset.

Dataset(s): We focus on relational classification datasets with high dimensionality ($d \gg 10$) to emphasize the speedup from parallelization of the GA for feature selection. We will use datasets from the UC Irvine Machine Learning Repository [1] shown in table 1.

Dataset	Dimension (d)	Samples (N)	Classes
TUNADROMD	241	4465	2
DARWIN	451	174	2
Toxicity	1203	171	2
TCGA Kidney Cancer	60660	1024	3

Table 1: Datasets found from [1]

Methodology: A high level overview of the GA is shown in figure 1. We denote the chromosome as a binary vector indicating which features to include (1) and exclude (0) for the ML model training, where a gene is a specific feature. The fitness value is the ML model's performance (such as cross entropy, F1-score, or any classification metric) on a validation split. The cross over and mutation GA operators will randomly "mix" and "flip" the binary vectors via binary operations between the ML model's chromosomes with the highest fitness to create a new population of chromosomes. This process is iterative repeated until user-specific convergence criteria is met.

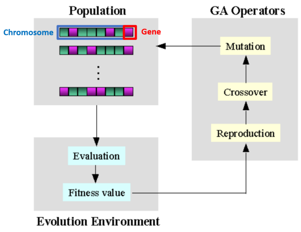

Figure 1: General Genetic Algorithm Flow Diagram [3]

References

[1] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: `http://archive.ics.uci.edu/ml`.

[2] Riccardo Leardi, Riccardo Boggia, and M Terrile. “Genetic algorithms as a strategy for feature selection”. In: *Journal of chemometrics* 6.5 (1992), pp. 267–281.

[3] Ying-Hong Liao and Chuen-Tsai Sun. 2001. URL: `https://www.ewh.ieee.org/soc/es/May2001/14/Begin.htm`.

[4] Feng Tan et al. “A genetic algorithm-based method for feature subset selection”. In: *Soft Computing* 12 (2008), pp. 111–120.