Strategy for OCFS manuscript: (Nov 21, 2014)

-Feature selection background

1. State-of-the-art methods
2. Swarm robotics (collective decision making, social behavior/leadership,etc)
3. Agent-based modeling
4. PSO and binary PSO (Escalante 2009, Chen 2008, other relevant articles)

-Novel features:

1. Fitness function: based on CV-BER, CV-AUC (permuted and original)
2. Sequential backward/forward selection to find the optimal number of features
3. Behavior based binary PSO
4. Combined framework: optimal collaborative feature selection

A novel wrapper feature selection algorithm where feature subsets are generated based on a binary behavior based particle swarm optimization algorithm. The fitness of the subsets is evaluated based on the k-fold cross-validated balanced error rate (BER) using a support vector machine based predictive model.

PSO is an optimization technique based on the swarm intelligence. A novel binary PSO is introduced here where the particles follow different topologies depending upon their current behavior. Each particle is assigned a behavior {confusion, follow neighborhood best, follows global best, follow personal best}.

**Global optima**

One of the limitations of stochastic search algorithms is that they get stuck in local optima. The position and behavior of the particles in the swarm is reinitialized if the fitness does not change for x number of iterations.

**Fitness function**

The key in fitness evaluation is that the objective function is based on the difference between the mean balanced error rate between the cross-validated and permuted balanced error rate. This allows to account for imbalanced study designs where samples are unevenly distributed between different sample groups/classes (Reference). A second term in the fitness function adds a small penalty based on the number of features in the model. For instance, if two subsets of features have same predictive accuracy, the subset with fewer number of efatures will be selected based on the Ockhams razor theory (Reference). This allows to identify the subset with least number of features.

**Behavior based topology**

A) CMA rank based feature selection combined with backward or forward feature selection

1. Rank features based on different methods
2. Use a forward or backward selection scheme to find the best set of features (explain; accuracy vs # of features graphs)
3. List includes both univariate and multivariate methods. A) Univariate
4. methods=c("rf","rfe","limma","lasso","elasticnet","kruskal.test","f.test", "f.test", "elasticnet", "wilcox.test", "welch.test") -Two sample Welch t-test
5. -Wilcoxon rank-sum test
6. -F-test
7. -Kruskal-wallis test  -Limma (moderated t-test and moderated F-test)  -SAM (Significance analysis of microarrays)   B) Multivariate  -Random forest variable importance measure  -Recursive feature selection with Support Vector Machines  -Lasso penalized feature selection  -Elastic net based penalized feature selection  -Multiple Adaptive Regression Splines -Sparse Partial Least Squares Regression combined with penalized feature selection

B) B3PSO

Novel features in B3PSO:

1. Behavior-driven topology selection
   1. fully-connected if everyone is aware of global;
   2. scale-free topology
   3. star topology
   4. One way to evaluate is to see if this prevents getting stuck in a local optima is to see how many times the global fitness gets updated
   5. Markov decision processes? does it help?
2. Behavior updated after x iterations: does it help?
3. Number of iterations determined by the performance (CV-BER/AUC) threshold and no-change criteria
   1. Cite Escalante PSO where they mention that it is hard to determine the number of iterations a priori
4. Fitness function: takes into account both actual CV-BER and permuted CV-BER: does it help?
5. Parallel fitness function evaluation
6. kd-trees based nearest neighbor calculations
7. Options for Bootsrap/CV/AUC/BER
8. * + - * Plot the saturation point for different scenarios
         * Balancing exploration and exploitation

-OCFS description

1. Multi-stage feature selection framework
2. Stage 1 restrict search space to features selected by one or more feature selection methods
3. Stage 2: find the optimal set using the

-Comparison of different FS methods on a series of datasets

1. Use the datasets in Escalante2009 and the binary PSO papers

-Conclusion

**-Evaluation**

**a) Different topologies**

**b) Different methods**

**c) Different combinations**

**d) Different number of particles, iterations**

**e) Which topology converges early w.r.t fitness?**

Code location:

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Rcode/versionnov2014/

/home/stu/kuppal3/Research/Feature\_selection/Rcode/versionnov2014

Location of Datasets:

a) ARCENE

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/ARCENE/arcene\_train.data

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/ARCENE/arcene\_train.labels

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/ARCENE/arcene\_valid.data

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/ARCENE/arcene\_valid.labels

b) SRBCT\_Khan

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/SRBCT\_Khan/Khan.xtrain.txt

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/SRBCT\_Khan/Khan.ytrain.txt

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/SRBCT\_Khan/khan.xtest.txt

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/SRBCT\_Khan/khan.ytest.txt

c) DOROTHEA

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/DOROTHEA/dorothea\_train.data.txt

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/DOROTHEA/dorothea\_train.labels.txt

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/DOROTHEA/dorothea\_valid.data.txt

/Users/karanuppal/Documents/Gatech/Projects/Algorithms/TwostagePSO/Datasets/DOROTHEA/dorothea\_valid.labels.txt