MIDTERM PROJECT
FOR ADVANCED OPTIMIZATION METHODS
SPRING SEMESTER 2022

In this midterm project, we try to develop a nature-inspired metaheuristic optimization approach to select important biomarkers in a medical diagnosis problem.

Background: Medical diagnosis is an important step for determining a parent's status on a particular disease. Traditionally, doctors perform diagnosis based on their experience and expert knowledge, and it leads to some inaccurate determination when the symptoms are similar. With the aid of advanced technologies, more precise measures via medical instruments and methods like NGS help to improve the accuracy of the diagnosis. Recently with the rise of data science, results from big data analytics and deep learning techniques further provide essential recommendations to doctors. Even with all these new techniques and methods, the existence of the curse of dimensionality never disappears, i.e. the number of potentially important biomarkers is always much larger than the number of patients in most cases.

Goal: The ultimate goal of this midterm project is to develop a metaheuristic optimization method that can be practically used to identify important biomarkers that have significant impacts towards the patients' disease category.

Training Data: You can download the training data from the class website (data.xls). It is an Excel file with 4240 data points, each with 19 target biomarker values ($F01$-$F19$), 5 reference biomarker values ($R01$-$R05$), and 5 diagnosis results ($C01$-$C05$). About the diagnosis results, $C01$-$C05$ are five disease categories. Any entry in the columns of Cxx with labeled 1 indicates that the patient is diagnosed as Cxx category. If all five entries in these five columns are labeled 0, the patient is considered as healthy (experimental contrast).

Data Preprocessing: We do not use those 19+5 columns directly to build the diagnosis model, but the ratio among all target biomarkers to all reference biomarkers. This means for each patient, we will have 95 biomarker ratios ($F01/R01$, $F01/R02$, …, $F01/R05$, $F02/R01$, …, $F02/R05$, …, …, $F19/R05$) and we relabel them as x_1, x_2, …, x_5, x_6, …, x_{10}, …, …, x_{95}. Furthermore, we believe there exists some interactions being important among all these 95 biomarker ratios, and to make it simple, we only consider all 4465 two-way ratios ($x_i x_j$) and all 138415 three-way ratios ($x_i x_j x_k$) for $i, j, k \in [1, 95]$ and $i \neq j \neq k$. This means we have a total of 142975 biomarker ratios.

Testing Data: You can download two testing data from the class website (Test1.xls and Test2.xls). The same data preprocessing is required. Test1.xls is the testing data for you to build the best diagnosis model. Test2.xls is a set of dat without diagnosis results. After you decide the method (among 5 of your choice) is the best method, you need to use your method to determine the patients' status in each row of Test2.xls.

Details and Specific Tasks:

The input of the optimization should be a list of biomarker ratios. You can pick no more than 30 ratios among all 142975 choices in each input.

To verify the goodness of an input, you need to use the training data to build a logistic regression model on each disease category, i.e. you should have five models with binary responses.

Then you use those five models to predict the disease categories of the patients in the testing data (Test1.xls). Finally you compare your predicted categories with the true categories. At this stage, you should have five accuracy values associated to five categories. Take the average of five values as the objective function value. It is obvious that the large this average accuracy, the better the input.

Depending on your choices on the optimization methods, there should be some parameter settings. Choose the reasonable values for these parameters. The only two common setups for all of you include: (1) a maximum of 100 particles for a population-based method (and run accordingly for a single-state method for comparison), (2) a maximum of 1000 iterations.

After all simulations, you should be able to pick the best method out of your five choices for this biomarker diagnosis problem. Then use your method to predict the diagnosis status of all 987 patients in Test2.xls.

Here is a list of tasks you need to complete in this midterm project:

1. You need to write down the formal statement of optimization of this problem. Define the input X, objective function $f(X)$, and constraints if exists.
2. Pick five optimization methods (starting from simulated annealing) and write the program codes for the biomarker diagnosis problem.
3. For each method, use the training data (Data.xls) and the testing data (Test1.xls) to logistic regression models with no more than 30 ratios.
4. Write down the best subsets of biomarker ratios for each method and their average accuracy after the simulation is finished.
5. Check if their accuracy are different by statistical tests and draw the progress diagrams for all five methods.
6. Pick the best method among five. Then predict all 987 patient status in Test2.xls.
7. Summarize the strength and weakness of the five methods of your choice.

What to hand in:

1. A report that includes detailed descriptions of the tasks and the results.
2. The program codes you used to obtain the results in your report.
3. A complete Test2.xls with predicted values (1 or 0) in the columns $C01$-$C05$.

Scores:
Let me emphasize again here: I do not care if you get all answers correct or not, but I care the process. I look into your report and see how you finish all tasks with correct approach and appropriate settings. I also look into your codes to see if there are any incorrect parts. You will get a high score if your concepts are correct, and your codes are runable without big problems and conceptual mistakes. Since all methods you can choose are randomized methods, some lucks are needed for you to get all answers correct (It is actually quite impossible).

Due Date: 2022/04/20