

**SCHOOL OF ENGINEERING AND COMPUTER SCIENCE**

**Assignment Cover Sheet**

**Full Name: \_\_\_\_\_\_\_\_\_\_\_\_\_Tao Shi\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_**

**Student ID: \_\_300409943\_\_ \_\_Course: \_\_\_\_\_ COM307 \_ \_\_\_**

**Assignment No.: \_\_\_\_\_2\_\_\_ \_ Due Date: 08 May at 11:59pm**

**Part 1: Neural networks for Classification**

**1. Determine and report the network architecture, including the number of input nodes, the number of output nodes, t****he number of hidden nodes (assume only one hidden layer is used here). Describe the rationale of your choice.**

**Answer:**

The number of input nodes is 4, because there are four numeric attributes.

The number of output nodes is 3, because there are three classes, where each class refers to a type of iris plant.

The number of hidden nodes is 2. Because the best is to have as few nodes as possible, firstly make the guess of 2, by using the neural network simulator of BPNN, I found the training is successful basically, after then I set the number of hidden nodes as 1, the training is unsuccessful. Consequently, I decided the number of hidden nodes is 2 eventually.

**2. Determine the learning parameters, including** **the learning rate, momentum,** **initial weight ranges, and any other parameters you used. Describe the rationale of your choice.**

**Answer:**

The learning rate is 0.2. In practical training process of BPNN, 0.2 is good starting point, when I set the learning rate higher, the accuracies of training data and test data decrease.

Initially, I set the momentum as 0, then I change it to 0.5, but there are no changes of results.

The initial weights range from -1 to 1. But when I change the range to [-0.5 0.5], there are no changes of results.

In addition, the number of epochs could be very large, but the user could presses CTRL-C to stop the training and causes the weights to be saved, which can be used to test the data sets later.

**3. Determine your** **network training termination criteria. Describe the rationale of your decision.**

**Answer:**

The network training termination criteria include critical error (mce in BPNN), and the percent of classification accuracy, I set them as 0.01 and 101% respectively, which means even though the percent of classification accuracy has reached to 100%, the critical error must be less than 0.01. The practical result illustrates the termination criteria will improve the accuracy.

**4. Report your results (average results of 10 independent experiment runs) on both the training set and the test set. Analyse your results and make your conclusions.**

**Answer:**

The average accuracy on training set is 85.2%, and the average accuracy on test set is 86.8%.

RUN1

Number of incorrect classifications: 11/75

Number of incorrect classifications: 10/75

RUN2

Number of incorrect classifications: 11/75

Number of incorrect classifications: 10/75

RUN3

Number of incorrect classifications: 11/75

Number of incorrect classifications: 10/75

RUN4

Number of incorrect classifications: 11/75

Number of incorrect classifications: 8/75

RUN5

Number of incorrect classifications: 11/75

Number of incorrect classifications: 9/75

RUN6

Number of incorrect classifications: 12/75

Number of incorrect classifications: 11/75

RUN7

Number of incorrect classifications: 11/75

Number of incorrect classifications: 11/75

RUN8

Number of incorrect classifications: 11/75

Number of incorrect classifications: 10/75

RUN9

Number of incorrect classifications: 11/75

Number of incorrect classifications: 10/75

RUN10

Number of incorrect classifications: 11/75

Number of incorrect classifications: 10/75

For more detailed results (result of training process, final weights and mse), please find the file in the folder of Assignment2\_programs/part1.

**5. (optional/bonus) Compare the performance of this method (****neural networks) and the nearest neighbour methods.**

**Answer:**

I suppose the performance of nearest neighbour methods is better than neural networks according to this case, especially when k = 3, the classification accuracy of KNN is 96.00%. Furthermore, the nearest neighbour methods are straightforward and the principle is obvious compared with the black box of neural networks.

However the performance of different methods is problem dependent. In some cases, neural networks are very powerful, especially for nonlinear regression, the method can represent any function with enough hidden nodes.

**Part 2: Genetic Programming for Symbolic Regression**

**1. Determine a good terminal set for this task.**

**Answer:**

The package of ECJ is applied in this problem.

I suppose the terminal set is x and random number r (function X and RegERC in ECJ). However, if I only set x as terminal set, a nice program is still offered by the GP, which means the random number does not play a key role.

**2. Determine a good function set for this task.**

**Answer:**

Because the problem is not complicated, I only set {+ , -, \*, %} as the function set, in my program the relevant functions are {Add, Sub, Mul, Div}.

**3. Construct a** **good** **fitness function and describe it using plain language (and mathematical formula, or other formats you think appropriate).**

**Answer:**

The fitness function can be described as the mean squared error (MSE), the error in the problem is the difference between the calculated by the program evolved by GP and output variable y according to each input variable x given in the question. The mathematical formula is as follows:

The less the result above is, the better the performance of the program is. I modified the KozaFitness.java in ECJ: if the error is less than 0.00001, it is regarded as 0.

And if is less than 0.01, the is regarded as the same as the output variable y according to the input variable x, and the number of the situation is labeled as Hits in code.

The greater value of Hits is, the better the performance of the program is.

**4. Describe the relevant parameter values and the** **stopping criteria you used.**

**Answer:**

The most default parameter values of ECJ are applied in the problem except the rates of reproduction, crossover and mutation:

Population = 1024. Generations = 51, tournament size = 7, ramped half-and-half method.

In term of max depth (program size): the crossover and mutation are 17, grow for subtree mutation is 5 (as many as the minimal depth), and the halfbuilder is 6(minimal depth is 2 by the way).

And reproduction rate is set as 5%, crossover rate: 90%, mutation rate: 5%.

In the problem, the stopping criteria is at generation 51 or the standardized error is 0, and the outfile offers the Best Individual of Run.

**5. List three different best programs evolved by GP and the fitness value of them (you need to run your GP system several times and report the best programs of the runs).**

**Answer:**

ECJ can print out trees in a somewhat C-style format. It is done by the command like: java ec.Evolve -file tutorial4.params -p gp.tree.print-style=dot. The three different best programs evolved by GP and fitness value are as follows.

**Program 1:**

((x \* x) + (((x + x) + (x % x)) % (x - x))) - ((x \* ((x + x) - (x \* x))) \* x)

The fitness value: 0, Hits=20.

**Program 2:**

(((x + x) \* (x + x)) % (x - x)) - (((x + x) \* (x \* x)) - (((x % x) + (x \* x)) \* (x \* x)))

The fitness value: 0, Hits=20.

**Program 3:**

((x % x) + (x \* x)) + (((x \* x) - (x + x)) \* (x \* x))

The fitness value: 0, Hits=20.

**6. (Optional, bonus) Analyse one of the best programs to reveal why it can solve the problem in the task.**

**Answer:**

According to one best program: ((x % x) + (x \* x)) + (((x \* x) - (x + x)) \* (x \* x)) and its simplified form:, we can learn about the relationship between the input variables and output variables, even though we do not know the distribution of data set, the independence of the input variables, even any statistical background knowledge. This is the main problem in the task, also is the most important advantage of GP method.

**Part 3: Genetic Programming for Classification**

**1. Determine a good terminal set for this task.**

**Answer:**

The package of ECJ is applied in this problem.

The terminal set in this task includes 9 numerically valued attributes and random number r (function X1 to X9 and RegERC in ECJ). However, regarding to the program evolved by GP, not all attributes play roles, even sometimes constants do not appear in the results.

**2. Determine a good function set for this task.**

**Answer:**

Because the problem is complicated, I set {+ , -, \*, %, ABS, EXP, LOG, SIN, COS,} as the function set, in my program the relevant functions are {Add, Sub, Mul, Div, Abs, Exp, Log, Sin, Cos}.

**3. Construct a good fitness function and describe it using plain language (and mathematical formula, or other formats you think appropriate).**

**Answer:**

Firstly I set the value of class as 1 or -1, when the classification of instances is benign or malignant respectively.

The fitness function can be described as the mean squared error (MSE), the error in the problem is the difference between the calculated by the program evolved by GP and according to each instance with 9 numerically valued attributes given in the question. The mathematical formula is as follows:

The less the result above is, the better the performance of the program is.

In addition, if the class of the instance is benign, and if as well, which means the classification is correct, the number of the situation is labeled as Hits in code, otherwise the classification is incorrect.

If the class of the instance is malignant, and if as well, which means the classification is correct, the number of the situation is labeled as Hits in code, otherwise the classification is incorrect.

The greater value of Hits is, the better the performance of the program is.

**4. Describe the relevant parameter values and the stopping criteria you used.**

**Answer:**

The most default parameter values of ECJ are applied in the problem except the rates of reproduction, crossover and mutation:

Population = 1024. Generations = 51, tournament size = 7, ramped half-and-half method.

In term of max depth (program size): the crossover and mutation are 17, grow for subtree mutation is 5 (as many as the minimal depth), and the halfbuilder is 6(minimal depth is 2 by the way).

And reproduction rate is set as 5%, crossover rate: 90%, mutation rate: 5%.

In the problem, the stopping criteria is at generation 51 or the standardized error is 0, but the outfile offers the Best Individual of Run.

**5. Split the original data set into a** **training set training.txt and a test set test.txt with appropriate format for the GP package you have chosen. Describe your main considerations of the splitting.**

**Answer:**

Because the volume of training set should be more than volume of test set, and the 70:30 conventions. I split the original 699 instances set into training set including 499 instances and a test set including 200 instances.

**6. Report the classification accuracy (average accuracy over 10 independent experiment runs) on both the training set and the test set.**

**Answer:**

The average accuracy on training set is 96.97%, and the average accuracy on test set is 98.5%.

RUN1

Training set:

Fitness: Standardized=0.13408752413321123 Adjusted=0.8817661588899897 Hits=477

Test set:

Fitness: Standardized=0.06388460458864684 Adjusted=0.9399515658812002 Hits=197

RUN2

Training set:

Fitness: Standardized=0.0815470364033396 Adjusted=0.9246014887392023 Hits=489

Test set:

Fitness: Standardized=0.02545718310943427 Adjusted=0.9751747966382741 Hits=199

RUN3

Training set:

Fitness: Standardized=0.0917302978252641 Adjusted=0.9159771437982516 Hits=485

Test set:

Fitness: Standardized=0.0690530506387867 Adjusted=0.9354072741315075 Hits=195

RUN4

Training set:

Fitness: Standardized=0.08155805397756549 Adjusted=0.9245920700441131 Hits=486

Test set:

Fitness: Standardized=0.060320966448576624 Adjusted=0.9431106538894399 Hits=197

RUN5

Training set:

Fitness: Standardized=0.12023540323824415 Adjusted=0.8926695202716484 Hits=480

Test set:

Fitness: Standardized=0.06002499939863871 Adjusted=0.9433739775640277 Hits=197

RUN6

Training set:

Fitness: Standardized=0.1009057851434958 Adjusted=0.9083429422342953 Hits=485

Test set:

Fitness: Standardized=0.033962492634063986 Adjusted=0.9671530709517875 Hits=197

RUN7

Training set:

Fitness: Standardized=0.12759429278187662 Adjusted=0.8868437933761707 Hits=482

Test set:

Fitness: Standardized=0.0678433447721557 Adjusted=0.936466949853369 Hits=196

RUN8

Training set:

Fitness: Standardized=0.1135528369414493 Adjusted=0.8980265388633553 Hits=482

Test set:

Fitness: Standardized=0.029214384537188543 Adjusted=0.9716148695780952 Hits=198

RUN9

Training set:

Fitness: Standardized=0.09790720902227946 Adjusted=0.9108237852728294 Hits=485

Test set:

Fitness: Standardized=0.047050047377221 Adjusted=0.9550641848543173 Hits=197

RUN10

Training set:

Fitness: Standardized=0.08482303682811919 Adjusted=0.9218093329985593 Hits=488

Test set:

Fitness: Standardized=0.0396851499796092 Adjusted=0.9618296462343552 Hits=197

**7. List three best programs evolved by GP and the fitness value of them.**

**Answer:**

**Program 1:**

sin(sin(sin(sin(sin(cos(rlog(x1)) + cos(rlog(x6 + rlog(rlog(x1))))) + cos(rlog(x3 + rlog(x1)))) + cos(rlog((x2 \* x7) + rlog(x2)))) + sin(sin(cos(x6 + rlog(rlog(x6 + (x6 + x5)) + ((rlog(x6 + x5) \* (x6 + x5)) - x1))) + (cos(rlog(x6 + rlog(x8))) + cos(rlog(rlog(x3) \* ((rlog(x6 + x5) \* (x8 + x5)) - x1))))) + sin(sin(cos(abs(rlog(x5 \* (x6 + x8)))) + (sin(cos(rlog(x1)) + cos(x6 + rlog(x8))) + cos(rlog(x3 + rlog(x1))))) + cos(rlog(abs(cos(rlog(x3 + x6)) \* rlog(x1)) + x6))))) + sin(sin(cos(x6 + rlog(rlog((rlog(x6 + x5) \* (x6 + x5)) - x1) + ((rlog(cos(x6 + rlog(x8))) \* (x6 + x5)) - x1))) + (cos(rlog(x6 + rlog(x8))) + cos(rlog(rlog(x3) \* ((rlog(x6 + x5) \* (x8 + x5)) - x1))))) + sin(sin(cos(abs(abs(rlog(rlog(x2) \* x1)))) + (cos(rlog(x3 + x6)) \* rlog(x1))) + cos(rlog(abs(cos(rlog(x3 + x6)) \* rlog(x1)) + x6)))))

Training set: Fitness value: 0.0815470364033396, Hits=489

Test set: Fitness value: 0.02545718310943427, Hits=199

**Program 2:**

sin(sin(sin(cos(rlog(rlog(x8 \* x3) \* x3)) + cos(rlog(x5 % cos(rlog(x6))))) + cos(rlog(x6 % rlog(exp(cos(0.19327037658223234)) - rlog(exp(cos(x6 % x2)) - rlog(x6 % x2)))))) + sin(cos(rlog(x6 % cos(rlog(sin(sin(cos(sin((x8 \* x7) \* x3)))) % rlog(x1))))) + cos(rlog(x1))))

Training set: Fitness value: 0.1009057851434958, Hits=485

Test set: Fitness value: 0.033962492634063986, Hits=197

**Program 3:**

sin(abs(rlog((abs(rlog(abs(rlog(sin(cos(x6)) - (x6 + rlog(x4 + x1)))) - (x6 + x2))) + rlog(abs(x2) + (sin(x8) + cos(cos(rlog(abs(x2))))))) - (x6 + x2))) + rlog((abs(rlog((abs(rlog(x4 + x1)) + (x6 + x2)) - ((x4 + x1) + x2))) + rlog((abs(x2) + cos(rlog(((x6 + x2) + rlog(x6 + rlog(x4 + x1))) + x1))) + (abs(x2) + cos(rlog(((abs(rlog(x4 + x1)) + (x6 + x2)) + rlog(x6 + x6)) + x1))))) + x1))

Training set: Fitness value: 0.1135528369414493, Hits=482

Test set: Fitness value: 0.029214384537188543, Hits=198

**8. (optional, bonus) Analyse one of best programs to reveal why it can solve the problem in the task.**

**Answer:**

According to one best program as follows:

y = sin(sin(sin(sin(sin(cos(rlog(x1)) + cos(rlog(x6 + rlog(rlog(x1))))) + cos(rlog(x3 + rlog(x1)))) + cos(rlog((x2 \* x7) + rlog(x2)))) + sin(sin(cos(x6 + rlog(rlog(x6 + (x6 + x5)) + ((rlog(x6 + x5) \* (x6 + x5)) - x1))) + (cos(rlog(x6 + rlog(x8))) + cos(rlog(rlog(x3) \* ((rlog(x6 + x5) \* (x8 + x5)) - x1))))) + sin(sin(cos(abs(rlog(x5 \* (x6 + x8)))) + (sin(cos(rlog(x1)) + cos(x6 + rlog(x8))) + cos(rlog(x3 + rlog(x1))))) + cos(rlog(abs(cos(rlog(x3 + x6)) \* rlog(x1)) + x6))))) + sin(sin(cos(x6 + rlog(rlog((rlog(x6 + x5) \* (x6 + x5)) - x1) + ((rlog(cos(x6 + rlog(x8))) \* (x6 + x5)) - x1))) + (cos(rlog(x6 + rlog(x8))) + cos(rlog(rlog(x3) \* ((rlog(x6 + x5) \* (x8 + x5)) - x1))))) + sin(sin(cos(abs(abs(rlog(rlog(x2) \* x1)))) + (cos(rlog(x3 + x6)) \* rlog(x1))) + cos(rlog(abs(cos(rlog(x3 + x6)) \* rlog(x1)) + x6)))))

we can find the relationship between the property of breast cancer and almost all of the nine attributes, such as Clump Thickness(CT, x1 in program), Uniformity of Cell Size (USz, x2 in program), Uniformity of Cell Shape (UShp, x3 in program), Marginal Adhesion (MA, x4 in program), SingleEpithelial Cell Size (SESz, x5 in program), Bare Nuclei (BN, x6 in program), Bland Chromatin (BC, x7 in program), Normal Nucleoli (NN, x8 in program) and Mitoses (M, x9 in program). By the program and values of the nine attributes of a patient, the result calculated could help doctors to diagnose the breast cancer.