Third International Competition: Protecting rivers and streams by monitoring chemical concentrations and algae communities solved with the use of GAdC

Dirk Devogelaere, Marcel Rijckaert
K.U.Leuven – Chemical Engineering Department
De Croylaan 46, 3001 Heverlee
Belgium
Dirk.devogelaere@cit.kuleuven.ac.be
Mark J. Embrechts
Rensselaer Polytechnic Institute – Dept. Environmental and Energy Engineering, Troy, NY 12180, USA

ABSTRACT: This paper presents a solution to the competition for protecting rivers and streams by monitoring chemical concentrations and algae. In the introduction the problem is stated. The next part presents a brief overview of the Genetic Algorithm driven Clustering method (GAdC) used to tackle this problem. Then the different steps to approach this problem are described.

1. Introduction (problem statement)

During the research study water quality samples were taken form sites on different European rivers of a period of approximately one year. These samples were analyzed for various chemical substances. In parallel, algae samples were collected to determinate the algae population distributions. While the chemical analysis is cheap and easily automated, the biological part involves microscopic examination, requires trained manpower and is therefore both expensive and slow.

The relationship between the chemical and biological features is complex and can be expected to need the application of advanced techniques. Typical of such real-life problems, the particular data set for the problem contains a mixture of qualitative, linguistic and numerical measurement values, with much of the data being incomplete.
The competition task is the prediction of algae frequency distributions on the basis of the measured concentrations of the chemical substances and the global information concerning the season when the sample was taken, the river size and its velocity. The two last variables are given as linguistic variables.

340 data sets were taken and each contains 18 values. The first 11 values of each data set are the season, the river size, the fluid velocity and 8 chemical concentrations that might be relevant for the algae population distribution. The last 7 values of each data set (AG1 … AG7) are the distribution of different kinds of algae. These 7 kinds are only a very small part of the whole community, but for the competition we limited the number to 7. The value 0.0 means that the frequency is very low. The data set also contains some empty fields that are labeled with the string XXXXX.
Each participant in the competition received 200 complete data sets (training data) and 140 data sets (evaluation data) containing only the 11 values of the river descriptions and the chemical concentrations.
This training data is to be used in obtaining a 'model' providing a prediction of the algae distributions associated with the evaluation data.

2 Basic Concept of the method GAdC

This section presents the Genetic Algorithm driven Clustering (GAdC) algorithm developed by the authors and designed for regression analysis. This algorithm consists of two parts: the training part where the cluster centra and scalars are determined and a test part where the model is evaluated towards unknown test data. In the training part a steady-state GA as described by De Jong was utilized. It uses overlapping populations with a user-specifiable amount of

overlap. Fundamentally it contains all the parts of a genetic algorithm. The next subsections discuss several aspects of the design of GAdC.

2.1 Genetic Algorithm driven Clustering

The basic idea of GAdC is to use a genetic algorithm where the traditional clustering performance measure is used as the cost function for evaluating the goodness of the clustering scheme. The clustering performance measure contains the traditional cluster distance (i.e., the sum of the Euclidean distances of all the cluster members from the cluster center for all the clusters) and several penalty functions to allow the algorithm to change the number of active clusters while the GA evolves and to discourage clusters with less than a specified number of members (typically 1). A variable number of clusters is an inherent part of this algorithm and is actually achieved by starting from a relatively large number of clusters (typically 10) and to encourage the presence of empty clusters (i.e., clusters with no members) by adding a negative penalty to the cost function. The GA just guesses the cluster centers for the number of clusters specified, but several of the guessed clusters will be empty.

2.2 Chromosome representation

For any genetic algorithm (GA), a chromosome representation is needed to describe each individual in the population of interest. In the GAdC it consists of floating point numbers with values within the variables upper and lower bounds. The variables are the cluster centra (one for each feature in each cluster) and the scalars (one for each feature).

2.3 Selection function

The selection of individuals to produce successive generations plays an important role in a genetic algorithm. A probabilistic selection is performed based upon the individual's fitness such that the better individuals have an increased chance of being selected. Here we use linear scaling in combination with roulette wheel selection.

2.4 Genetic operators

Genetic operators provide the basic search mechanism of the GA. The operators are used to create new solutions based on existing solutions in the population. There are two basic types of operators: crossover and mutation. In GAdC uniform crossover and flip mutation are applied.

2.5 Initialization and termination

The GA must be provided with an initial population. The common method of randomly generating solutions for the entire population within the solution space was applied for the GAdC.
The GA moves from generation to generation selecting and reproducing parents until a termination criterion is met. Each generation the algorithm creates a temporary population of individuals, adds these to the previous population, then removes the worst individuals in order to return the population to its original size. The most frequently used stopping criterion is a specified maximum number of generations as we use in GAdC.

2.6 Evaluation function

The evaluation function is the driving force behind the GA. The evaluation function is called from the GA to determine the fitness of each individual generated during the search. The evaluation function consists of the sum of four terms: 1) a cluster distance cost function, 2) a misclassification penalty, 3) an empty_cluster penalty, 4) a 1_element_cluster penalty.

3 Tackling the problem

3.1 Empty fields in training data

The data lines 62 and 199 are not used due to the high amount of XXXX in it. The missing value in column D is replaced by the average. Missing data of columns E, F, J and K are calculated by concerning those columns as targets.

3.2 Empty fields in evaluation data

The empty field in column D is replaced by the average of the column. A model is built to predict the values in the empty fields of columns F, J and K.

3.3 Prediction of the Algae frequency in the evaluation data

The information about the seasons is not used in the regression. The data about the river size and the fluid velocity is replaced by the values 0.0; 0.5 and 1.0 depending on the indication small, medium or high. First of all a model was built to predict the column AG1. This data is also used to predict the value of column AG2. In this way the values of all the columns are predicted.
The results are presented in a txt file and presented as given in table 1:

Table 1. Subset of the results as presented in the txt file

AG1	AG2	AG3	AG4	AG5	AG6	AG7
3.9	17.7	7.3	0.4	6.6	5.3	2.0
9.9	5.5	3.5	1.5	9.7	8.3	1.9
4.5	9.8	8.8	2.7	4.2	4.6	5.3
12.4	5.4	1.7	1.0	12.9	12.8	2.3
16.7	2.7	5.4	0.7	2.1	2.9	2.1
26.3	3.9	0.8	2.6	2.2	2.3	1.3
53.2	3.4	6.6	1.1	0.9	0.6	0.2
35.6	3.9	2.8	1.5	2.0	0.8	2.2
29.5	2.3	7.1	1.0	2.0	2.1	2.3
24.0	3.5	3.7	1.8	5.3	5.1	2.5
1.2	0.3	0.9	10.1	2.0	46.9	1.1
1.7	1.3	2.0	18.5	3.5	0.9	2.0
37.0	1.8	5.5	0.9	1.9	1.4	1.0
25.9	1.6	1.8	2.4	1.1	0.3	1.2
26.8	1.5	1.7	1.8	0.8	0.3	1.1
16.6	3.0	4.0	0.3	4.6	6.3	2.1
2.7	15.5	7.4	4.9	5.6	3.0	1.8
2.5	16.0	1.6	0.4	1.4	7.0	2.3
34.5	1.5	2.9	3.0	0.7	0.2	1.4
54.1	1.6	1.0	3.2	0.0	0.0	0.0
3.3	20.0	4.9	1.3	3.5	5.4	0.8
5.8	18.6	5.9	1.3	4.6	3.8	1.1
7.4	13.3	2.0	1.7	16.8	21.7	1.4