Optimising the Mutual Information of Ecological Data Clusters using Genetic Algorithms

Holger R. Maier[a], Lucy Radbone[a], Tom Finkemeyer[a], Tiana Hume[a], Miranda Butchart[a] and Peter Goonan[b]

[a]Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, The University of Adelaide. Email: hmaier@civeng.adelaide.edu.au

[b]Environment Protection Authority, Adelaide, South Australia.

Abstract: The Australian River Assessment System (AusRivAS) is a nation-wide program designed to assess the health of Australian rivers and streams. The general AusRivAS method involves the establishment of a database of reference sites, which are sites that are considered to be minimally affected by anthropogenic impacts. These sites are then grouped into clusters of similar macroinvertebrate communities. The clusters are analysed to find relationships between the physical, geographical and chemical properties of sites in a cluster and the corresponding macroinvertebrate communities. The relationships found are then used to predict the macroinvertebrate communities at non-reference sites that would be expected if these sites were equivalent to least disturbed reference conditions. To determine the level of river health, the expected macroinvertebrate community is compared with the observed community. As part of AusRivAS, the clustering step is conducted using the statistical Unweighted Pair Group Arithmetic Averaging (UPGMA) method. A potential shortcoming of this approach is that it uses a linear performance measure for grouping similar data points. A recently developed approach for clustering ecological data (MIR-max) overcomes this limitation by using mutual information as the performance measure. In this paper, an alternative to the MIR-max technique (MIRA4) is proposed, which uses genetic algorithms for optimising the overall mutual information of the ecological data clusters. The MIR-max and MIRA4 approaches are applied to the South Australian combined season riffle AusRivAS data, and the results obtained are compared with those obtained using the UPGMA method. The results indicate that the overall mutual information values of the clusters obtained using MIR-max and MIRA4 are significantly higher than those obtained using the UPGMA method, and that the use of genetic algorithms is successful in determining clusters with higher overall mutual information values compared with those obtained using MIR-max for the case study considered.

Keywords: *AusRivAS; River health assessment; Clustering; Genetic algorithm; Mutual information*

1. BACKGROUND

1.1. Introduction

The complexity of river ecology makes any assessment of river health difficult. Traditional assessment measures, including physical and chemical parameters, may fail to recognize significant changes in river health, such as the effect of acute pollution events. In order to overcome these limitations, biological indicators are being used increasingly. Such indicators include macroinvertebrates, fish, algae, diatoms, microorganisms and macrophytes. Macroinvertebrate data, in particular, have been used extensively for the assessment of river health, as they are present in almost all rivers, and different species have various sensitivities to environmental stress. In addition, they only travel within a limited range and have a lifespan that is adequate for detecting disturbances, while sufficiently brief to detect recolonisation after a disturbance (Dallas, 2000).

The River Invertebrate Prediction and Classification Scheme (RIVPACS), which is used to assess river health in Britain, was the first regional-scale model that incorporated the use of macroinvertebrate data as an alternative to physical and chemical data. Since the development of RIVPACS, similar methods of assessing river health have been set up in other countries, including the Australian River Assessment System (AusRivAS) and a similar system in California (Hawkins et al., 2000). Recently, O'Connor & Walley (2002) developed a River Pollution Diagnostic System (RPDS) for

the British Environment Agency using a novel information-theoretic clustering system called MIR-max designed specifically for the purpose (Walley and O'Connor, 2001). An interesting feature of RPDS is that it does not rely on the use of reference sites.

1.2. Australian River Assessment System

The Australian River Assessment System (AusRivAS) was established in 1992 as a nation-wide program designed to assess the health of Australian rivers and streams. The general AusRivAS method involves the establishment of a database of reference sites, which are sites that are considered to be minimally affected by anthropogenic impacts. These sites are then grouped into clusters of similar macroinvertebrate communities. The clusters are analysed to find relationships between the physical, geographical and chemical properties of sites in a cluster and the corresponding macroinvertebrate communities. The relationships found are then used to predict the macroinvertebrate communities at non-reference sites that would be expected if these sites were equivalent to least disturbed reference conditions. To determine the level of river health, the expected macroinvertebrate community is compared with the observed community.

1.3. Clustering of Reference Sites

As part of AusRivAS, the clustering of references sites is performed using the statistical Unweighted Pair Group Arithmetic Averaging (UPGMA) method (Davies, 1994), which is an agglomerative hierarchical technique (Kaufman and Rousseeuw, 1990). Sites are agglomerated in a stepwise fashion to produce a hierarchical order, which is presented in the form of a dendrogram (Kaufman and Rousseeuw, 1990). As part of the agglomeration process, Euclidean distance is used as the performance measure to assess the similarity of sites based on the macroinvertebrate communities present.

A limitation of agglomerative methods, such as UPGMA, is that if clusters are joined suboptimally, they can never be separated. Thus, errors created in previous steps of the clustering process cannot be overcome (Kaufman and Rousseeuw, 1990). Another potential shortcoming of the UPGMA method is that it uses a linear performance measure (i.e. Euclidean distance) for grouping similar data points, which can fail to capture the non-linear relationships that are a feature of ecological systems.

The MIR-max system introduced by Walley and O'Connor (2001) overcomes the limitations of the UPGMA approach outlined above. MIR-max clusters the data by maximising the mutual information (MI) between the clusters and the attributes of the data, and then arranges the clusters in a two-dimensional space in a way that aims to preserve their relative positions in *n*-dimensional data space. The mutual information criterion (Fraser and Swinney, 1986) is used as the performance measure as it caters for both linear and non-linear dependence between variables. In addition, the approach is not agglomerative, enabling data points to move between clusters during the clustering process. In this paper, only the clustering aspect of MIR-max is addressed.

As part of the MIR-max approach, mutual information is maximised using a hill-climbing approach. This involves selecting two sampling sites from different clusters and swapping them. If the MI score is increased as a result of the swap, the change takes place; if not, the sites return to their original clusters. This process is continued until no improvement is made for a user-defined number of iterations (Walley and O'Connor, 2001). In this paper, an alternative approach for maximising mutual information is introduced (MIRA4), which uses genetic algorithms as the optimisation engine.

1.4. Objectives of Research

The objectives of this research are:

1. To develop an alternative approach for clustering ecological data based on mutual information (MIRA4) by replacing the hill-climbing approach for optimising mutual information currently used in MIR-max with genetic algorithms (GAs). GAs are robust, stochastic search algorithms that are based on Darwin's theory of natural selection. In recent years, GAs have been shown to have advantages over classical optimisation methods (Goldberg, 1989) and have become one of the most widely used techniques for solving a number of hydrology and water resources problems (Vasquez et al., 2000).
2. To compare the MI between clusters and the attributes of the data points to be clustered obtained using the UPGMA, MIR-max (use of hill-climbing to optimise MI) and MIRA4 (use of GAs to optimise MI) clustering approaches for the South Australian combined season riffle AusRivAS data.

2. PROPOSED APPROACH

The proposed approach for clustering ecological data involves the use of mutual information as the performance measure for determining the similarity between clusters and the attributes of the data points to be clustered, and to maximise the performance measure (i.e. mutual information) using genetic algorithms. Details of the proposed approach are given below.

2.1. Performance Measure

The mutual information between two given variables X and Y is given by (Sharma, 2000):

$$MI = \iint f_{X,Y}(x,y) \log_e \left[\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \right] dxdy \quad (1)$$

where $f_X(x)$ and $f_Y(y)$ are the marginal probability density functions of X and Y respectively, and $f_{X,Y}(x,y)$ is the joint probability density function of X and Y. The mutual information measures the reduction in uncertainty of Y as a result of knowledge about X. If there is no dependence between X and Y, then the two random variables are statistically independent and, by definition, the joint probability density $f_{X,Y}(x,y)$ would equal the product of the marginal densities ($f_X(x)\, f_Y(y)$). If this is the case, the MI score would be zero. If, on the other hand, the random variables were strongly related, the value of MI would be high.

When clustering ecological data, the objective is to determine to which of n clusters each data point should belong, such that the mutual information between the cluster (C) and the m attributes (X) of the samples is maximised (Walley and O'Connor, 2001). If it is assumed that each of the attributes X_j occurs in one of s states (k = 1 to s), then the mutual information M(C, X_j) is given by (Walley and O'Connor, 2001):

$$M(C, X_j) = \sum_{i=1}^{n} \sum_{k=1}^{s} \alpha_{ijk} \log_e \left[\frac{\alpha_{ijk}}{\beta_i \gamma_{jk}} \right] \quad (2)$$

where α_{ijk} = probability of finding attribute X_j in its kth state in cluster C_i, β_i = prior probability of class C_i, γ_{jk} = prior probability of finding attribute X_j in its kth state and

$$\alpha_{ijk} = \frac{\rho_{ijk}}{T} \quad (3)$$

$$\beta_i = \frac{q_i}{T} \quad (4)$$

$$\gamma_{jk} = \frac{r_{jk}}{T} \quad (5)$$

where p_{ijk} = number of samples in cluster C_i with attribute X_j in its kth state, q_i = number of samples in category C_i, r_{jk} = number of samples with attribute X_j in its kth state and T = total number of samples. The total mutual information (G) between clusters and the attributes of the data points to be clustered is given by (Walley and O'Connor, 2001):

$$G = \sum_{j=1}^{m} M(C, X_j) \quad (6)$$

In the case of AusRivAS, a reference site is considered a sample, and (2) measures the dependence between the cluster allocation of a particular reference site (C) and the state (s) of a specific macroinvertebrate community (X) at that reference site. The state of a macroinvertebrate community either refers to presence or absence (i.e. s = 2), or one of a number of discrete abundance levels (i.e. s = total number of discrete abundance levels). A high MI score signifies a high dependence between the cluster allocation and the state of the macroinvertebrate community. Consequently, sites that have macroinvertebrate communities in similar states will cluster together. By summing the mutual information score for each macroinvertebrate community sampled (see (6)), the overall mutual information for that cluster set can be determined.

2.2. Optimisation of Performance Measure

GAs are heuristic iterative search techniques that attempt to find the best solution in a given decision space based on a search algorithm that mimics Darwinian evolution and survival of the fittest in a natural environment (Goldberg, 1989). In keeping with genetics terminology, the decision space is referred to as the environment, the potential solutions to the optimisation problem are called chromosomes (or strings of information that represent a decision set) and the total number of chromosomes is called the population size. The iterations of the optimisation process are called generations and the GA proceeds by evaluating the best sets of chromosomes in the population at each generation. These sets of chromosomes are found by evaluating the objective function for each chromosome in the population and by using this objective function value to indicate the fitness of the chromosomes. The chromosomes in a population compete with each other for survival, based on their fitness levels, and more fit individuals are given a higher probability of mating and reproducing and hence influencing the following generations. Through competition for survival, the population evolves to contain high-performing chromosomes.

An advantage GAs have over traditional optimisation techniques is that they do not require the use of the gradient of a fitness function, only the value of the fitness function itself. Another advantage of GAs is that they search from a population of points, investigating several areas of the search space simultaneously, while traditional optimisation methods only test one scenario at a time. GAs therefore have a greater chance of finding the global optimum. For a more detailed description of GAs, the reader is referred to Goldberg (1989).

In order to apply GAs to maximising the mutual information between clusters and the attributes of the data points to be clustered, the problem has to be formulated as follows.

1. The decision variables are the cluster allocations of each of the data samples (e.g. reference sites). Consequently, the number of decision variables is equal to the number of data samples, and the number of values each decision variable can take is equal to the number of clusters to which the data samples should be allocated.
2. A solution consists of cluster allocations for all data points. Each solution is represented as a string of integers (i.e. a chromosome), as shown in Figure 1. The total number of integers is equal to the number of data points (e.g. reference sites), and the values each integer can take range from 1 to n, where n is the number of clusters the data points can be allocated to (Figure 1).

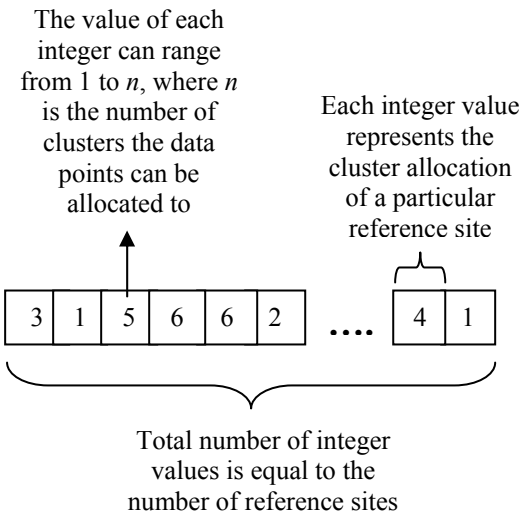

Figure 1. Representation of a solution as a string of integers (chromosome)

The optimisation process is summarised in Figure 2. At the start of the process, a number (population) of solutions are generated at random, and the "fitness" of each solution is calculated in accordance with (6). Next, the "fittest" chromosomes are selected as potential parents for the next generation. In this research, tournament selection was used, where two chromosomes from the population are paired off at random, and the "fitter" of the two chromosomes survives, whereas the other chromosome is eliminated. In order to ensure that the population size stays constant, the number of tournaments conducted is equal to the population size. In this research, the two chromosomes that participate in each tournament were chosen from the total population pool at random, with replacement.

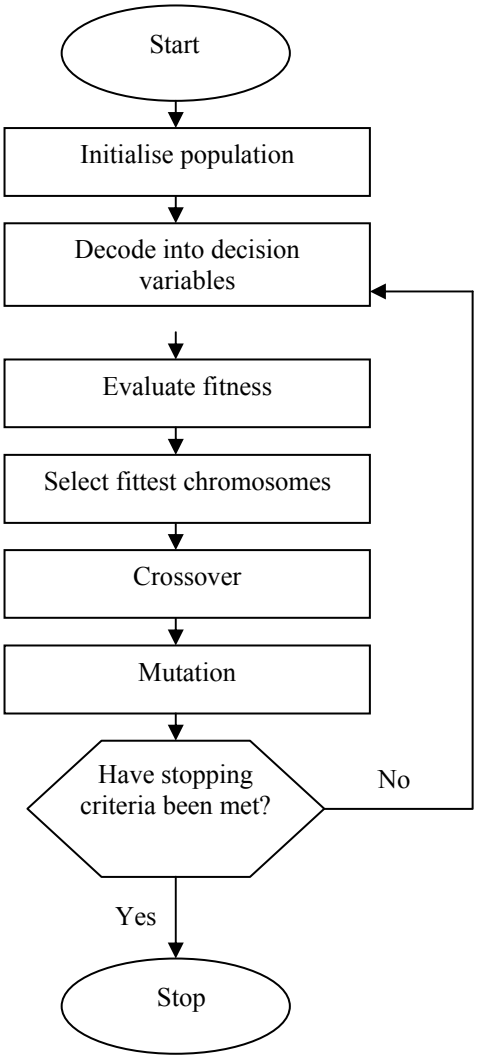

Figure 2. Steps in the genetic algorithm optimisation process

Next, members of the parent pool, which consist of the winners of the tournaments, are paired up at

random and have the opportunity to exchange information via a process called crossover. The probability that a pair of strings will exchange information is referred to as the probability of crossover. In this research, two point crossover was used, in which two parent chromosomes are "cut" at two identical, random locations, and the integers in the parent chromosomes (i.e. cluster locations) between the cuts are swapped.

In order to ensure sufficient exploration of the decision space, the value of some of the integers in a chromosome (i.e. cluster locations) are changed at random in a process called mutation. Whether mutation of a particular integer occurs is governed by the probability of mutation.

The chromosomes obtained after the application of the processes of selection, crossover and mutation (i.e. the children) become the parents in the next generation and the process is repeated until certain stopping criteria, such as the completion of a fixed number of iterations, are met. In this research, elitism was employed, which ensures that the fittest member of a generation is guaranteed to survive the selection process in the next generation. This ensures that there is no reduction in fitness from one generation to the next.

3. CASE STUDY

3.1. Data

The data used in the analyses are the South Australian combined (Autumn/Spring) riffle macroinvertebrate data, which have been collected by the South Australian Environment Protection Authority (EPA). The data contain 151 reference sites (i.e. data points) and information on 67 macroinvertebrate families (i.e. $m = 67$). In order to compare the results directly with those obtained using the UPGMA method currently used in the AusRivAS model, only presence / absence data were considered (i.e. $s = 2$) and the number of clusters used was 6 (i.e. $n = 6$).

3.2. Analyses Conducted

In order to meet the objectives of this research, the following analyses were conducted:

1. The mutual information between the cluster allocation of a particular reference site and the state (i.e. presence / absence) of a specific macroinvertebrate community at that site was calculated for the clusters obtained from the UPGMA method using (2). The overall mutual information was then calculated using (6).
2. The available data were clustered using the MIR-max software and the mutual information between the cluster allocation of a particular reference site and the state (i.e. presence / absence) of a specific macroinvertebrate community at that site was calculated for the clusters obtained using (2). MIR-max clustering was continued until there was no further change in the objective function value for a large number of iterations. The overall mutual information was then calculated using (6).
3. The available data were clustered using the proposed approach (MIRA4) and the mutual information between the cluster allocation of a particular reference site and the state (i.e. presence / absence) of a specific macroinvertebrate community at that site was calculated for the clusters obtained using (2). The overall mutual information was then calculated using (6). The software code required for implementing the MIRA4 approach was developed in Fortran 77.

   In relation to the GA, a population size of 30 and a stopping criterion of 5,000 generations were used for all simulations. The optimal probabilities of crossover and mutation were determined by trial and error. The probabilities of crossover ranged from 0.6 to 1.0. The best results obtained were for a probability of crossover of 0.9, although the performance of the GA was relatively insensitive to this parameter. Probabilities of mutation investigated ranged from 0 (i.e. no mutation) to 0.1. The best results were obtained when the probability of mutation was 0.001. The performance of the GA decreased significantly for higher values of probability of mutation, such as 0.01.

4. RESULTS AND DISCUSSION

The overall mutual information values obtained using the three clustering approaches investigated are shown in Table 1. It can be seen that, for the case study considered, the overall mutual information values of the MIR-max (11.84) and MIRA4 (12.03) approaches were significantly higher than that obtained using the UPGMA approach (9.14). However, as pointed out by Walley and O'Connor (2001), it is not surprising that clustering methods that are designed to optimise mutual information achieve higher MI scores than clustering methods that use an alternative performance measure.

The results in Table 1 also suggest that, for the case study considered, the genetic algorithm approach proposed as part of MIRA4 is more

successful than the hill-climbing approach for optimising the mutual information between clusters and the attributes of the samples to be clustered used as part of MIR-max. However, whilst the GA results obtained are encouraging, it is recognised that further comparative studies are needed on more challenging case studies, such as that used by O'Connor and Walley (2002) to develop RPDS, which clustered 6038 samples into 250 clusters. In addition, the robustness of the two approaches to different starting positions in objective function space needs to be investigated. Finally, it should be noted that the MIR-max approach is still under development, and recent trials with an optimisation approach similar to simulated annealing have resulted in mutual information values that are approximately 2% higher on average than those obtained using the hill-climbing approach currently used.

Table 1. Overall mutual information obtained using the three clustering methods investigated

Clustering method	Overall mutual information
UPGMA	9.14
MIR-max (hill climbing)	11.84
MIRA4 (GA)	12.03

5. CONCLUSIONS

A new approach for clustering ecological data was introduced in this paper, which uses mutual information as the performance measure and genetic algorithms for optimising this performance measure. The approach was applied to the South Australian combined season riffle AusRivAS data.

It was found that the overall mutual information between clusters and the attributes of the samples to be clustered obtained using the new approach (MIRA4) was significantly higher than that obtained using the UPGMA clustering method (12.03 compared with 9.14), which is currently used in the South Australian AusRivAS model, and slightly higher than that obtained using the MIR-max approach (12.03 compared with 11.84), which uses a hill-climbing approach for optimising mutual information. This indicates that the proposed approach shows promise, but further comparative tests are needed.

6. ACKNOWLEDGMENTS

The authors would like to thank Mark O'Connor from the Centre for Intelligent Environmental Systems at Staffordshire University for his valuable advice and comments on this manuscript and supplying the MIR-max software.

7. REFERENCES

Dallas, H. F., The Derivation of Ecological Reference Conditions For Riverine Macroinvertebrates, Southern Waters Ecological Research and Consulting, 2000.

Davies, P. E., River Bioassessment Manual, National River Processes and Management Program, Monitoring River Health Initiative, 1994.

Fraser, A. M. and Swinney, H. L., Independent coordinates for strange attractors from mutual information, *Physics Review A*, 33(2), 1134-1140, 1986.

Goldberg, D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 412, Reading, MA, 1989.

Hawkins, C. P., Norris, R. H., Houge, J. N. and Feminella, J. W., Development and evaluation of predictive models for measuring the biological integrity of streams, *Ecological Applications*, 10(5), 1456-1477, 2000.

Kaufman, L. and Rousseeuw, P. J., *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley and Sons, Inc., Brisbane, 1990.

O'Connor M. A. & Walley W. J., River Pollution Diagnostic System (RPDS) – computer-based analysis and visualisation for bio-monitoring data. *Water Science and Technology*, 46(3), 17-23, 2002.

Sharma, A., Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification, *Journal of Hydrology*, 239, 232-239, 2000.

Vasquez, J. A., Maier, H. R., Lence, B. J., Tolson, B. A. and Foschi, R. O., Achieving water quality system reliability using genetic algorithms, *Journal of Environmental Engineering, ASCE*, 126(10), 954-962, 2000.

Walley, W. J. and O'Connor, M. A., Unsupervised pattern recognition for the interpretation of ecological data, *Ecological Modelling*, 146, 219-230, 2001.