FuREAP: A fuzzy-rough estimator of algae population

by

Qiang Shen, Alexios Chouchoulas
FuREAP: A fuzzy-rough estimator of algae population

Qiang Shen, Alexios Chouchoulas
Informatics Research Report EDI-INF-RR-0124
DIVISION of INFORMATICS
Centre for Intelligent Systems and their Applications
May 2001


Abstract:
Concern for environmental issues has increased in recent years. Waste production influences humanity’s future. The alga, an ubiquitous single-celled plant, can thrive on industrial waste, to the detriment of water clarity and human activities. To avoid this, biologists need to isolate the chemical parameters of these rapid population fluctuations.

This paper proposes a Fuzzy-Rough Estimator of Algae Populations (FuREAP), a hybrid system involving Fuzzy Set and Rough Set theories that estimates the size of algae populations given certain water characteristics. Through dimensionality reduction, FuREAP significantly reduces computer time and space requirements. Also, it decreases the cost of obtaining measurements and increases runtime efficiency, making the system more viable economically. By retaining only information required for the estimation task, FuREAP offers higher accuracy than conventional rule induction systems. Finally, FuREAP does not alter the domain semantics, making the distilled knowledge human-readable.

The paper addresses the problem domain, architecture and modus operandi of FuREAP, and provides and discusses detailed experimental results.

Keywords: Algae population estimation, fuzzy logic, rough sets, rule induction, dimensionality reduction

Copyright © 2002 by The University of Edinburgh. All Rights Reserved

The authors and the University of Edinburgh retain the right to reproduce and publish this paper for non-commercial purposes.

Permission is granted for this report to be reproduced by others for non-commercial purposes as long as this copyright notice is reprinted in full in any reproduction. Applications to make other use of the material should be addressed in the first instance to Copyright Permissions, Division of Informatics, The University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland.
FuREAP: A Fuzzy-Rough Estimator of Algae Populations

Qiang Shen\textsuperscript{1} Alexios Chouchoulas\textsuperscript{2}

\textit{Institute for Representation and Reasoning, Division of Informatics, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, UK}

\begin{abstract}
Concern for environmental issues has increased in recent years. Waste production influences humanity’s future. The alga, an ubiquitous single-celled plant, can thrive on industrial waste, to the detriment of water clarity and human activities. To avoid this, biologists need to isolate the chemical parameters of these rapid population fluctuations. This paper proposes a Fuzzy-Rough Estimator of Algae Populations (FuREAP), a hybrid system involving Fuzzy Set and Rough Set theories that estimates the size of algal populations given certain water characteristics. Through dimensionality reduction, FuREAP significantly reduces computer time and space requirements. Also, it decreases the cost of obtaining measurements and increases runtime efficiency, making the system more viable economically. By retaining only information required for the estimation task, FuREAP offers higher accuracy than conventional rule induction systems. Finally, FuREAP does not alter the domain semantics, making the distilled knowledge human-readable. The paper addresses the problem domain, architecture and modus operandi of FuREAP, and provides and discusses detailed experimental results.
\end{abstract}
Key words:
Algae population estimation; fuzzy logic; rough sets; rule induction;
dimensionality reduction.

1 Introduction

Concern for environmental issues has increased greatly in the last decade. The production of waste, toxic and otherwise, from a vast number of different manufacturing processes and industrial plants is one of the most important issues. It influences directly the future of humanity’s food and water supply. As such, extreme care has to be taken in order to maintain the balance. To wit, it has become clear that even changes in farming and sewage water treatment can affect the chemistry and ecology of rivers, lakes and even the sea.

In particular, growing algae\(^3\) communities are detrimental to water clarity, while complex water life like fish can also be endangered, due to changes in the oxygen content of the water. There can even be effects on human activities in such areas, since toxic effects may be present in relation to algae growth. There is a multitude of different species of alga, each with their own characteristics. Most of them respond very rapidly to changes in their environment. Ecologies where algae are present are thus heavily dependent on adequate chemical and

\(^1\) qiangs@dai.ed.ac.uk
\(^2\) alexios@dai.ed.ac.uk
\(^3\) The alga is a single-celled plant that has, over a period of three and a half billion years, evolved into the most successful coloniser of almost any known ecology on the planet.
physical balance in the environment. Measuring and reducing the impact that farming, manufacturing and waste disposal have on nutrient content in rivers has, thus, attracted much attention.

To help in this task, an intelligent tool would be highly desirable. The system should locate the parameters that control algae population fluctuations and use this information to estimate these changes. Such a system could aid in a number of areas, including simulating hypothetical scenarios and predicting trends in algae communities, in addition to its intended estimation task. The tool would greatly simplify the collection of chemical concentration measurements in the water by isolating the absolutely necessary tests that need to be conducted. This would simplify the testing process itself. Having such an intelligent system would also decentralise the entire process, allowing individual testers to take samples and obtain results in situ. This would in turn reduce the cost associated with these measurements, and minimise response times.

In building an intelligent system to cope with the complexity of the estimation task and the interrelated nature of the chemicals diluted in the water, however, knowledge acquisition could be an obstacle. Eliciting knowledge from any source of data is notorious for its difficulty. Whether the source of information is a human expert, or a dataset of experimental measurements, extracting general knowledge from it is a serious bottleneck in the development of a knowledge-based system in general.

The complexity of the domain typically adds a lot of difficulty to knowledge acquisition. Data collected for complicated, real-life application domains is likely to have a large number of attributes, many of which are superfluous in
performing the desired task — yet their presence has detrimental effects on all knowledge acquisition and machine learning software. Such effects range from drastic training speed reduction to rendering knowledge acquisition intractable for the domain. The runtime of the system is affected as well. Having additional quantities to measure tends to be a slow, error-prone and cost-ineffective situation. Also, unknown, unavailable or inaccurate values cannot be ruled out, while instruments inevitably develop faults of their own, and the humans reading them are not infallible. The absence of experts to interpret and check the data often proves to be another problem.

This paper discusses a Fuzzy-Rough Estimator of Algae Populations (FuREAP), a system that determines the population of several species of river algae based on physical and chemical water measurements.

The aim of FuREAP is to induce low-dimensionality rulesets from historical data pertaining to the distribution of algae as a function of the above measurements. FuREAP is a hybrid system comprising a dimensionality reduction subsystem based on Rough Set theory [12] and a fuzzy Rule Induction Algorithm (RIA) [10] to extract information from datasets. The Rough Set Attribute Reduction (RSAR) subsystem (described later on in this paper) works as a pre-processor to the RIA.

There are several existing methodologies relevant to the task at hand, both from the point of view of applications and computational approaches. Au and Chan’s FAPACS algorithm [1] discovers fuzzy association rules in relational databases by locating pairs of attributes that satisfy an ‘interestingness’ measure that is defined in terms of an adjusted difference between the observed and
expected values of relations. Through the use of fuzzy rules, the algorithm is able to express linguistically the regularities and exceptions discovered within the data. Romahi and Shen [15] have applied this approach to Financial Forecasting with success. Hayashi et al [6] have documented modifications to the Fuzzy ID3 (itself an augmentation of Quinlan’s original ID3 [14]) rule induction algorithm to better support fuzzy learning and to increase the system’s learning accuracy. Janikow [8] has proposed modifications to decision trees to combine symbolic decision trees with approximate reasoning, offered by fuzzy representation. This approach redefines the methodology for knowledge inference, proposing different methodologies based on fuzzy control and rule-based systems. The result is a method best suited to relatively stationary problems, where readability and smooth output transitions are of the essence.

A disadvantage of these techniques and most similar ones is sensitivity to high dimensionality. This may be remedied using Principal Components Analysis (PCA) [7], a well-known tool for data analysis and transformation that is also known as Hotelling transform, Karhunen-Loève transform, or eigenvalue transform. PCA maximizes the variance of data points along the axes by locating a set of new axes and transforming the dataset. The new axes are constructed in order of decreasing variance, so that the first attribute in the new, transformed dataset has the most variance and the last attribute has the least variance. Correlations between attributes in the former sample space are removed in the new one, hence redundancies are also reduced. It is thus possible to reduce the dimensionally of a dataset by transforming it with PCA and then only using a certain number of attributes, starting with the first one.
However, although PCA is an efficient methodology, it irreversibly destroys the underlying semantics of the dataset. Further reasoning about the data is almost always humanly impossible, prohibiting the use of PCA as a dataset pre-processor for symbolic or fuzzy learning systems. By implication, only purely numerical (non-symbolic) datasets may be processed by PCA.

Unfortunately, semantics-preserving dimensionality reduction (feature selection) approaches tend to be domain specific, utilising well-known features of specific application domains. RSAR is an approach to feature that preserves the underlying semantics of the data while offering reasonable generality. The modularity of the framework is such that RSAR may be used with numerous different rule induction algorithms, in addition to the RIA.

The authors have previously [17] proposed an application of the RIA+RSAR techniques to different tasks, including monitoring the operation of a large urban water treatment plant [13]. The ever-increasing demand for dependable, trustworthy intelligent diagnostic and monitoring systems, as well as knowledge-based systems in general, has focused the attention of researchers on the knowledge-acquisition bottleneck. The task of gathering information and extracting general knowledge from it is well-known to be the most difficult part of creating a knowledge-based system for any engineering task. The RIA+RSAR methodology eases this soup task of knowledge gathering in an automatic, efficient and domain-independent way.

The rest of the paper is arranged as follows. It starts by discussing issues pertaining to the estimation of algae population fluctuation. The theoretical background of the two major components of FuREAP is then discussed. Ex-
periments are then discussed, and results are shown and analysed. The paper is then concluded and further work is pointed out.

2 The Application Domain

As stated previously, the aim of developing FuREAP is to estimate the concentration of various different types of river alga, based on a set of chemical concentrations and other parameters [3]. To build the knowledge base, training samples were taken from different European rivers over the period of one year. These samples were analysed to quantify the presence of several chemicals, including nitrates, nitrites and ammonia, phosphate, oxygen and chloride. The pH of the water was also measured. In addition, the algae population distributions for each of the species involved were determined in the samples. A number of additional factors were taken into account, such as the season, river size and flow rate.

It is relatively easy to locate relations between one or two of these quantities and a species of algae. However, the process involves expertise in chemistry and biology and requires well-trained personnel and microscopic examination that cannot be automated given the state of the art. Thus, the process becomes expensive and slow, even for a subset of the quantities involved here. There are complex relations at work between the attributes of this application domain, be they conditional or decision; algae may influence one another, as well as be influenced by the concentration of chemicals. As such, there is expected to be some redundancy in the data. An important reason for the present
Fig. 1. Density plots for three of the algae dataset attributes: two of the eight chemical concentration distributions (left and middle), and the distribution of population values for one species of alga.

development is utilising the RSAR technique.

The application domain requires FuREAP to be able to estimate the populations of seven different species of alga based on eleven attributes of the river sample:

- The time of year the sample was taken, given as a season,
- The size of the river,
- The flow rate of the water, and
- Eight chemical concentrations, including nitrogen in the form of nitrates, nitrites, ammonia, phosphate, the pH of the water, oxygen and chloride.

The dataset available for training includes 200 instances. The first three attributes of each instance (season, river size and flow rate) are represented as fuzzy linguistic variables. Chemical concentrations and algae population estimates are represented as continuous quantities. The dataset includes a few samples with missing values. Of the 200 instances, two exhibiting mostly unknown values were removed from the data because of their low quality.

To generalise given training samples, attributes of numerical values are preprocessed to become symbolic. As the first three conditional attributes are already represented in fuzzy terms, no such preprocessing is required for them.
Matters differ for the eight chemical concentrations. As with all concentrations, these exhibit an exponential distribution (as shown in figure 1). To ease processing, samples were converted to a logarithmic scale defined by $f(x) = \log(x + 1)$, where $x$ is the numerical measurement of an attribute\(^4\).

As can be expected, the distributions of the algae are also exponential. This, coupled with the fact that the decision attributes representing algae population counts are numerical, suggests the use of a similar treatment as above. The conditional attributes were thus transformed by $g(x) = \lfloor \log(x + 1) \rfloor$, where $x$ is the numerical measurement of the algae community’s population and $\lfloor \cdot \rfloor$ is the floor operator\(^5\). This quantisation is required because the proposed approach can only distinguish between discrete classes.

The seven decision attributes were thus quantised into four discrete values each to signify four increasing levels of abundance. This is reasonable because of the way the algae population ‘counts’ are obtained. It is assumed that the river’s water is perfectly homogeneous and that any sample of the water, no matter how small, is statistically representative. Water samples are thus obtained. A few drops of each sample are examined visually via microscope and the number of algae are counted. This allows for human errors in determining the population, as well as the fact that a number of drops of water from a sample of a river are not necessarily statistically representative of the entire river. Quantisation can help alleviate this problem, provided that the number

\(^4\) Concentrations are non-negative real numbers, hence it is necessary to add an arbitrary constant to avoid the logarithm of zero.

\(^5\) Yielding the maximum integer less than or equal to the floor operator’s operand.
Fig. 2. Block diagram of the entire system, including the training and runtime aspects.

of quantisation levels are chosen carefully. In addition, if the aim is to estimate the behaviour of algae communities, it is far more intuitive to provide linguistic descriptions like ‘normal’, ‘lower’ and ‘higher’ rather than estimated concentrations that have to be matched against tables and may again be subject to human error.

3 FuREAP in Detail

A block diagram of the system is shown in Fig. 2, showing both the training and runtime stages of the system. The diagram demonstrates both the identification and selection of significant attributes and the generation of the fuzzy ruleset.
During the training stage, seven unreduced datasets (one for each alga species) of eleven conditional attributes each are obtained from water samples (marked ‘observables’). The datasets are reduced with RSAR to obtain seven datasets of seven conditional attributes each. These are then provided to the RIA, which induces seven rulesets (one for each species of alga).

During the runtime stage, the water samples are analysed to obtain only seven of the original eleven conditional attributes (marked ‘reduced datasets’), as per the reduct set chosen by the RSAR. This simplifies, speeds up and reduces the costs associated with the data gathering stage. Along with the rulesets induced by the RIA in the training stage, these new, seven-attribute datasets are used by a fuzzy reasoner to provide the system’s user with estimations of the seven algae populations.

3.1 Inducing Inference Rules

The rule induction algorithm presented in [10] extracts fuzzy rules from real-valued examples. Although this data-driven RIA was proposed to be used in conjunction with neural network-based classifiers, it is independent of the type of classifier used [16]. Provided with training data, the RIA induces approximate relationships between the characteristics of the conditional attributes and those of the decision attributes. The conditional attributes of the induced rules are represented by fuzzy variables, facilitating the modelling of the inherent uncertainty of the application domain.

The algorithm generates a hyperplane of candidate fuzzy rules $(p_1 \wedge p_2 \wedge$
\( \cdots \land p_n \Rightarrow c \) by fuzzifying the entire dataset using all combinations of rule conditions. Thus, a domain with \( n \) conditional attributes, each of which is a fuzzy region fuzzified by \( f_x \) fuzzy sets (\( 1 \leq x \leq n \)), the hyperplane is fuzzified into \( \prod_{i=1}^{n} f_i \) \( n \)-dimensional clusters, each representing one vector of rule conditions. Each cluster \( p = \langle \mu_1, \mu_2, \ldots, \mu_n \rangle \) may lead to a fuzzy rule, provided that training examples support it. To obtain a measure of what classification applies to a cluster, fuzzy min-max composition is used. The conditional attribute values of each training example are fuzzified according to the fuzzy conditions \( \langle \mu_1, \mu_2, \ldots, \mu_n \rangle \) that make up cluster \( p \). For each example \( \underline{x} = \langle x_1, x_2, \ldots, x_n \rangle \), \( S_{\underline{x}, p}^c = \min (\mu_1(x_1), \mu_2(x_2), \ldots, \mu_n(x_n)) \) is calculated. This is the s-norm of example \( \underline{x} \) with respect to cluster \( p \) and classification \( c \). To give a measure of the applicability of a classification to cluster \( p \), the maximum of all s-norms with respect to \( p \) and \( c \) is calculated (this is dubbed a t-norm): 
\[
T_p^c = \max \{ S_{\underline{x}, p}^c \mid \underline{x} \in D_c \},
\]
where \( D_c \) is the set of all dataset examples that can be classified as \( c \). This is iterated over all possible classifications \( c \) to provide a full indication of how well each cluster applies to each classification.

A cluster generates at most one rule. The rule’s conditions are the cluster’s \( n \) co-ordinate fuzzy sets. The conclusion is the classification attached to the cluster. Since there may be t-norms for more than one classification, it is necessary to decide on one classification for each of the clusters. Such contradictions are resolved by using the uncertainty margin, \( \varepsilon \) (\( 0 \leq \varepsilon < 1 \)). This means that a t-norm assigns its classification on its cluster if and only if it is greater by at least \( \varepsilon \) than all other t-norms for that cluster. If this is not the case, the cluster is considered undecidable and no rule is generated. The uncertainty
margin introduces a trade-off to the rule generation process. In general, the higher \( \varepsilon \) is, the less rules are generated, but classification error may increase.

The RIA in use is NP-hard, and may become intractable when inducing rules for datasets with many conditional attributes \([2]\). The most important problem, in terms of both memory and runtime is dealing with the large numbers of combinations of fuzzy values. This is not so important when only a few attributes are involved. Applied to a more complex application, such as the present algae population estimator, without some means of attribute reduction, the algorithm’s intractable nature becomes evident, in terms of both time and space.

It is thus convenient and helpful to treat the creation of fuzzy-set vectors as the creation of a tree, as demonstrated in Fig. 3. In this context, a leaf node is one combination of membership functions, and each arc represents one evaluation of a membership function. The minimum membership is retained when creating the t-norms. Any membership function that evaluates to zero implies that all leaf nodes in the subtree will eventually evaluate to zero, too, because of the use of the \( \min(\cdot) \) function. A subtree is therefore useless and can be pruned if (and only if) its root node evaluates to zero. Figure 3 shows this
in the context for the XOR function applied to operands A and B, with fuzzy
sets representing positive, zero and negative values (P, Z and N respectively).
The leftmost evaluation of A yields zero and the algorithm does not evaluate
any further fuzzy membership values, effectively pruning the tree.

In an application domain where a reasonable degree of resolution is required,
it is not unusual to see quantities partitioned into five to seven fuzzy sets.
Assuming an average of six fuzzy sets per attribute and forty attributes, the
RIA would need to generate $6^{40}$ combinations. If the pruning algorithm is used
instead, and making the reasonable assumption that any value will belong to at
most two fuzzy sets, there is a worst case of $2^{40}$ combinations to be evaluated.
The time needed is nineteen orders of magnitude less than that needed for the
full tree traversal. The savings are significant, but the number of combinations
is still far too large.

3.2 Reduction of Attributes

In order to alleviate further the NP-hard nature of the RIA, the RSAR tech-
nique is employed. RSAR reduces redundancies by selecting those conditional
attributes that are most significant to the classification represented in the
dataset. At the same time, it ensures that no essential information is lost.
The approach makes use of conventional set theory operations and works by
maximising a quantity known as degree of dependency, $\gamma_P(C)$ [12].

The degree of dependency $\gamma_P(C)$ of a set $C$ of decision attributes with respect
to a set $P$ of conditional attributes provides a measure of how important $P$ is
in classifying the dataset examples into \( C \). If \( \gamma_P(C) = 0 \), then classification \( C \)
is independent of the attributes in \( P \), hence the conditional attributes are of
no use to this classification. If \( \gamma = 1 \), then \( C \) is completely dependent on \( P \),
hence these attributes are indispensable. Values \( 0 < \gamma_P(C) < 1 \) denote partial
dependency, which shows that only some of the attributes in \( P \) may be useful,
or that these attributes do not provide all the information required for the
classification task. Alternatively, partial dependency could indicate that the
training data is missing the required information altogether.

The naïve version of the RSAR algorithm evaluates \( \gamma_P(Q) \) for all possible
subsets of the dataset’s conditional attributes, stopping either when it either
reaches 1, or there are no more combinations to investigate. An improved
version, the QuickReduct algorithm [2], escapes the combinatorial, NP-hard
nature of the naïve version by searching the tree of attribute combinations in
a best-first manner. It starts off with an empty subset and adds attributes one
by one, each time selecting the attribute whose addition to the current subset
will offer the highest increase of \( \gamma_P(Q) \). The improved algorithm stops when
a \( \gamma_P(Q) \) of 1 is reached, or when all attributes have been added. Adding all
attributes may not necessarily result in a \( \gamma \) of 1, in which case the dataset
could not be correctly classified to begin with.

A second, updated version of the algorithm, QuickReduct II, has recently
been implemented. It terminates when a \( \gamma_P(Q) \) of 1 is reached, when all at-
tributes have been added, or when the addition of the last attribute to the
reduct set does not change \( \gamma \). This guarantees to obtain the smallest reduct.

RSAR does not compromise with a set of conditional attributes that contains
a large part of the information of the initial set — it attempts to reduce this attribute set without loss of information significant to the classification at hand. RSAR depends on the use of nominal attributes. However, this does not give rise to problems because the real attribute values are ultimately fuzzified for use by the RIA, becoming ordered symbolic values.

This fuzzification process [18] must split the domain into no fewer discrete values than needed, as this will otherwise result in too much loss of information, and hence in rulesets with reduced accuracy. Splicing the domain into more values than necessary will, on the other hand, tax the induction process, and is likely to result in less effective attribute reduction.

3.3 Application of Learned Rules

Fuzzy rules generated by the RIA resemble the following:

\[
\text{if river-size is large and flow-rate is slow then species-c is abundant.}
\]

It does not matter to the RIA what membership a dataset example’s river-size attribute has in the fuzzy set large, as long as its membership is higher by at least \( \varepsilon \) than the membership in any of the other sets covering the attribute.

During runtime operation, the system employs the induced rules to perform Min-Max fuzzy inference. For each datum to be classified, all rules are applied. Each condition of each rule is applied by evaluating the membership value of the corresponding attribute of the datum in the fuzzy set involved in the condition. For the example rule above, it would be necessary to calculate
the membership of the *river-size* attribute value in the *large* fuzzy set, and similarly for the membership of the current value of *flow-rate* in the *slow* set. The minimum of these membership values is then retained as a measure of how well the rule applies to the datum. The rule with the maximum such value is selected to provide the classification of the datum. If more than one rule has the maximum value and not all rules agree on the classification, the system declares its ignorance on the datum. It marks the datum as ‘undecidable’ and allows a more educated estimate to be made by the operator.

For efficient performance, the induced ruleset should be as small as possible. At the same time, classification accuracy should be optimal for the task at hand. High values of $\epsilon$ drastically reduce the size of the ruleset, but increase classification error. Low values behave in the opposite manner. These two factors have to be traded off to satisfy application-dependent specifications. For a given problem, a good choice for the $\epsilon$ that provides a balance between a resultant ruleset’s size and accuracy can be found by experiment, as illustrated below.

4 Experimental results

4.1 Presentation of Results

Experimental results are given as two types of graphs: estimation error and ruleset size. Both quantities are plotted against $\epsilon$, the uncertainty margin or tolerance which creates a trade-off between the estimation accuracy of the rule-
set and the number of learned rules it comprises. Note that estimation error, rather than estimation accuracy is shown here. This is done to emphasise the accuracy/size trade-off. Ruleset size grows exponentially, so graphs involving it are shown on a logarithmic scale. All seven algae species are shown separately on each graph as a family of curves. Also, please note that, in plotting the graphs, ‘undecidable’ answers by the RIA were considered wrong answers, thus giving slightly more conservative results.

For convenience, each of the seven alga species were processed separately by the RIA in order to provide seven different rulesets. Each ruleset models the behaviour of one species. The separate rulesets can be merged trivially, to form a single ruleset. Alternatively, the RIA can be applied to all seven to produce directly a single, unified ruleset. This latter choice is, of course, a more inelegant and inflexible solution than having separate algae models. Therefore, the following results are shown with respect to individual algae species.

4.2 The Results

It is, first of all, interesting to investigate what effects dimensionality reduction may have on the runtime performance of FuREAP. To show whether FuREAP has an impact on overall accuracy, the RIA algorithm was used to induce a ruleset from the entire, unreduced algae dataset [3]. The results are shown on the top row of figure 4. Then, FuREAP was instructed to reduce the dimensionality of the dataset and produce another ruleset from these reduced data. This resulted in a seven-attribute dataset selected from the original, eleven-attribute one. The results of testing the ruleset induced from this dataset are
Fig. 4. Algae estimation accuracy before (top) and after (bottom) dimensionality reduction. The left graphs show estimation error against the value of $\varepsilon$; the right graphs show ruleset size (in a logarithmic scale) against $\varepsilon$.

illustrated on the bottom row of figure 4. A typical fuzzy rule induced from the unreduced dataset looks like this:

If $Season$ is $Summer$ and $River-Size$ is $Small$ and $Flow-Speed$ is $Medium$ and $Concentration-1$ is $Medium$ and $Concentration-2$ is $Medium$ and $Concentration-3$ is $High$ and $Concentration-4$ is $High$ and $Concentration-5$ is $Low$ and $Concentration-6$ is $Medium$ and $Concentration-7$ is $Medium$ and $Concentration-8$ is $Medium$ then $Alga-Species-1$ is $Abundant$.

The same rule, induced from the reduced dataset, looks like this:

If $Season$ is $Summer$ and $Flow-Speed$ is $Medium$ and $Concentration-1$ is $Medium$ and $Concentration-4$ is $High$ and $Concentration-6$ is $Medium$ and $Concentration-
7 is Medium and Concentration-8 is Medium then Alga-Species-1 is Abundant.

The exact selected attributes were different for each alga species, although certain attributes were present in all seven reduct sets, namely the season (attribute one) and Concentrations 1, 4 and 7. The obtained reducts could not be verified based on empirical evidence because the dataset documentation mentions the names of the concentration attributes, but not their ordering in the data, hence it is needed to refer to the chemical concentrations by number rather than name. However, based on previous experience with RSAR [2; 17], it is expected that the reducts would overall make sense to an expert. It must also be noted, however, that it is difficult to verify directly the quality of selected attributes, in default of a suitable quality metric. The most accessible way is therefore to use the reduced and unreduced data to train a learning system, and compare the results. This gives an indirect measure of reduct quality.

There is a certain drop in accuracy (around 10%) after dimensionality reduction, which may indicate that the attribute reduction process has removed some of the necessary information. However, a full investigation of the domain reveals that inexpert fuzzification is largely responsible for the error during the rule-induction phase. The fuzzification of certain conditional attributes is less successful than others. This causes the removal of some of the better-fuzzified attributes during dimensionality reduction, leading to the observed drop in accuracy.

Despite this accuracy reduction, however, the rule set induced from the low-dimensionality data is around two orders of magnitude smaller than that gen-
erated from the unreduced dataset. Induction speed increases at a higher rate, making a strong argument for the use of FuREAP in applications where computer time and storage are at a premium. As stated previously, however, the speed and storage benefits are not limited to the training stage. They extend to the runtime use of the system. By reducing the dimensionality of the dataset, the arity of the rules is also decreased. This allows for fewer measured variables, which is important for dynamic systems where observables are often restricted, or where the cost of obtaining more measurements is high. In the river algae domain, for instance, providing different measurements has different costs attached. It is trivial to give the time of year and size of river, but flow rate may need extra equipment. Additionally, each of the measurements of concentration of chemicals may need its own process, requiring time, well-trained personnel and money. Reducing the number of measurements to be made significantly enhances the potential of the estimator system.

To show that the dimensionality reduction part of FuREAP performs as claimed, it is desirable to prove two further points: that the RSAR algorithm in FuREAP truly finds the smallest, best subset of conditional attributes of the dataset (known as a reduct); and that adding further attributes to this reduct does not produce better results.

To this end, two further experiments were conducted. In the first, numerous datasets of six attributes each were randomly generated from the original, eleven-attribute algae dataset. Rulesets were induced from these, and the average estimation error of all runs was plotted, as shown on the right graph of figure 5 (where the left graph is the reduced dataset error for Fig. 4, copied
Fig. 5. Comparison of estimation error after training on the reduct set of seven attributes (left), and random sets of six attributes (right).

Fig. 6. Comparison of estimation error after training on the reduct set of attributes (left), and the reduct set plus one random attribute (right).

here to ease comparison). Two empirical conclusions can be drawn from these results: first, not all attributes contribute the same information; second, the results obtained from random sets of attributes are worse than those obtained from the reduct set. The latter conclusion demonstrates that RSAR does indeed locate a relatively high-quality reduced attribute set.

In the second further experiment, the four remaining conditional attributes were added to the seven-attribute reduct one at a time. The aim was to show that more attributes do not necessarily imply higher accuracy. Rulesets were induced from these artificially produced attribute sets, and the results were
averaged. As shown on the right graph of figure 6 (again, the canonical, reduced results from Fig. 4 are shown on the left graph for comparison), error increased by adding an arbitrary attribute to the reduct. This leads to the conclusion that the reduct indeed leads to an accuracy loss that is acceptably low.

It is clear that FuREAP performs very well. This shows that real-world problems do contain a lot of redundancy which, once removed, allows highly accurate rule sets of low-arity rules to be induced. To reinforce the significance of the present approach, the performance of FuREAP is compared to that of a system employing rules generated using C4.5 [14] from the sample dataset. FuREAP is able to provide a estimation accuracy that surpasses that of C4.5, all the while using a smaller set of conditional attributes (as shown in table 1). Although C4.5 offers superior training speed, the number of attributes involved in the final system is very important, inasmuch as the cost, complexity and time requirements of obtaining each set of measurements is proportional to the number of measurements in each set.

5 Conclusion

Controlling and limiting waste production is a very important issue in preserving the fragile balance of river ecologies. River algae are very sensitive to changes in their environment, and in turn, can influence the well-being of more complex life forms. Ecologies where algae are present are thus heavily dependent on chemical and physical balance in the environment. Growth in
Table 1
Comparison between FuREAP and C4.5 with respect to accuracies and the number of conditional attributes involved.

<table>
<thead>
<tr>
<th>Algae</th>
<th>FuREAP</th>
<th>C4.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species</td>
<td>Error</td>
</tr>
<tr>
<td>a</td>
<td>15%</td>
<td>7</td>
</tr>
<tr>
<td>b</td>
<td>12%</td>
<td>7</td>
</tr>
<tr>
<td>c</td>
<td>16%</td>
<td>7</td>
</tr>
<tr>
<td>d</td>
<td>8%</td>
<td>7</td>
</tr>
<tr>
<td>e</td>
<td>11%</td>
<td>7</td>
</tr>
<tr>
<td>f</td>
<td>12%</td>
<td>7</td>
</tr>
<tr>
<td>g</td>
<td>9%</td>
<td>7</td>
</tr>
</tbody>
</table>

algae communities is associated with poor water clarity and various detrimental effects on other forms of river life, as well as humans. Thus, measuring and reducing the impact that farming, manufacturing and waste disposal have on river ecologies has attracted much attention recently, especially with respect to estimating and controlling river algae population fluctuations. Biologists are attempting to locate the chemical parameters that control the rapid growth of algae communities.

It would be desirable to have an intelligent computer-based system to locate these parameters and use this information to estimate population fluctuations. Such a system would serve in a number of ways: it would simplify the collection of measurements by isolating the absolutely necessary tests needed. It would also decentralise the entire process, allowing individual testers to sample rivers.
and obtain results rapidly and in situ. This would in turn reduce monetary
and time requirements.

This paper has described FuREAP, a Fuzzy-Rough Estimator of Algae Popu-
lations. The Approach integrates rough-set-assisted attribute reduction with
a potentially powerful rule-induction algorithm in a modular manner. The
RSAR subsystem helps reduce the dimensionality of the domain with which
the RIA subsystem has to cope. The RSAR algorithm has proved to be very
useful in stripping out insignificant information, while retaining more impor-
tant conditional attributes. The most desirable feature of this technique is the
fact that, unlike transformation-based approaches, it maintains the underlying
semantics of the dataset, enabling human experts to glean the distilled knowl-
edge. In this way the original RIA, which is sensitive to the dimensionality of
the dataset, becomes usable on datasets consisting of a moderately large num-
ber of attributes. The resulting learned ruleset becomes manageable, and may
even outperform rules learned using more conditional attributes. Reducing the
number of observables also decreases the cost of obtaining measurements and
increases runtime efficiency, making the system more viable economically.

FuREAP is, of course, not perfect. Much improvement work remains to be
done. For example, as indicated above, fuzzification plays a very important
role in obtaining high-quality rulesets, because of the design of the original
RIA. An optimising pre-processor for domain fuzzification would be very help-
ful. Techniques for fuzzy-set optimisation, typically by the use of a genetic
algorithm, to search for the most suitable fuzzy-set definitions, have been pro-
posed [5]. Such a technique can be added to the system in an effort to increase
the quality of fuzzification when no expert is readily available. Also, an investigation into rule set post-processors is required; a suitable post-processing module would increase rule set accuracy while decreasing its cardinality. Several initial attempts at this have recently been reported [11; 9; 4], providing a very useful starting point for such further work. Finally, to test and perhaps to reinforce the generality of the proposed approach, work is ongoing to apply the framework to the building of knowledge bases for non-physical (but still real) problem domains. It is also very interesting to investigate the actual effects of utilising an alternative fuzzy rule induction algorithm to the method currently employed, especially one that would allow simple associations between conditional and decision attributes to be created [15]. This may further boost both the efficiency of both training and runtime aspects of the system, in addition to the gains offered by attribute reduction.

References


[3] ERUDIT, European Network for Fuzzy Logic and Uncertainty Modelling in Information Technology. Protecting Rivers and Streams by Monitoring Chemical Concentrations and Algae Communities (3rd


