Case Libraries and Information Theoretic Case Matching for Soil and Water Resources Management

Sarah Dorner[1], Christopher Pal[1], Edwin Ongley[2], David A. Swayne[3]
[1] *Computing Research Laboratory for the Environment, University of Guelph, Guelph, Ontario N1G 2W1, Canada*
[2] *Emeritus Scientist, Environment Canada, P.O. Box 5050, Burlington, Ontario L7R 4A6, Canada*
[3] *Department of Computing & Information Science, University of Guelph, Guelph, Ontario N1G 2W1, Canada*

Key words: Decision theoretic expert systems, agricultural applications

Abstract: This paper presents an alternative or complementary technique to exhaustive watershed modelling, in which a case library of previous watershed studies is compiled. Watershed studies found within existing literature are indexed by a set of characteristic parameters or features believed to be most relevant for estimating sediment, nutrient and pesticide transport. Cases are indexed using parameters describing broad watershed features. The case library can then be queried for a potential new project. One can then search through the database of cases as one would a standard database. The cases can be organised into a decision tree in which various nodes of the tree represent tests of the parameter values. Measures of the information content of each parameter reflect its ability to predict observed transport measurements. Cases are retrieved that nearly match parameter estimations for a particular project under consideration. The existing cases that are found in the same leaf of the decision tree are then presented for review and analysis of the proposed project

1. INTRODUCTION

"Future development of sustainable agriculture in a water-scarce world requires that the off-site impacts on water quality be estimated at the time an agricultural project is proposed. This will ensure that any degradation of water quality due to agriculture can be anticipated and factored into basin-

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35503-0_29
R. Denzer et al. (eds.), *Environmental Software Systems*

wide, integrated water resource planning. The methodologies for estimating sediment, nutrient and pesticide runoff, such as modelling, are extremely limited in developing countries because of the absence of data, the expense of collecting reliable data, and the absence of relevant reference studies for model development and calibration." [Ongley et al 1997]

Watershed models are often constructed to evaluate the impact of agricultural projects on water quality. However, these types of models require a person with some modelling experience (even deep knowledge of the particular models), they have large data requirements and a substantial amount of time and effort is needed to calibrate the model to produce a valid result. The use of watershed models to evaluate the impact of potential agricultural projects on water quality is complicated by two main factors. First, it takes measurement data taken over a number of years to accurately calibrate a model. Thus, problems arise where decisions must be made in a time frame shorter than a calibration time frame. Secondly, the collection of data is costly and the decision-making agency may not have the resources to collect such data. Thus, as there are a number of problems associated with using models for management decisions, decision-makers may be interested in alternative or complementary methodologies for evaluating the impact of potential agricultural projects.

One such methodology involves searching though previous watershed studies for relevant similar cases. However, finding relevant similar cases may be a difficult task. A number of factors may complicate the process of finding relevant cases.

1. The literature may be widely distributed in publications, some of which may not be readily available.
2. The decision-maker may not be an expert in the area and thus the relevancy of similar cases may not be obvious.

Having a solid case library is vital for the case-matching approach to function effectively. The development of a case library is the single-most time-consuming task. A large number of cases is required to represent as many types of watersheds as possible within the scope of the project. The information must be collected in a meaningful way in order for the search strategy to provide reliable and representative matched cases. Another challenge to the development of a case library beyond the time requirements is that data collection methods may change over time, or different factors may be considered more or less important as more research is conducted. A case library should be updated and expanded as new information becomes available.

In order to facilitate the process of finding relevant cases a literature search was performed in which existing studies were indexed according to a number of broad scale attributes or parameters. Once the cases were

compiled and indexed, a database application was created in which the cases could be searched based on standard queries on these index parameters. The compilation and indexing of literature studies addresses the first critical factor discussed earlier by simply assembling existing studies into one information repository. However the second factor involves the problem of identifying similar cases. This problem was addressed by measuring the amount of information contained within the index parameters relevant to the classification of the discretized output parameters of the case. The technique is motivated by, relatively recent research in the Machine Learning and Statistics community leading to the development of algorithms for constructing Decision Trees or Classification and Regression Trees (CARTs) [Brieman 1984] based on information theoretic measures. Our technique differs from the more standard use of decision trees in that the tree is not used for classification of the case of interest. Rather, the construction of the decision tree is used to partition the case library into sets of "similar" cases, where "similar" is defined as: cases containing parameters characteristic of a particular range of output parameters.

2. METHODOLOGY

2.1 The Selection of Relevant Parameters

The case library for the prototype was narrowed to only include cases containing sediment and nutrient yield data from studies of dry-land, rain-fed agricultural watersheds of area less than 100 km^2. Although there are many factors contributing to soil loss and water pollution from agriculture, a set of key parameters describing broad-scale features was defined. Ideally, the set of important parameters would be defined solely by those factors that are the most important for describing a watershed and its processes, and each case would be reported within the defined framework. For the prototype, it was necessary to define a set of parameters based on the available data reported in the literature that matched as best as possible a template defined before-hand.

The parameters selected for the prototype corresponded to broad-scale descriptive parameters based on the Universal Soil Loss Equation that is used in many water quality models. Thus, there are parameters to describe climate, topography, soil type, and crop type and management practices. It was necessary to define a classification scheme that could be used to obtain a match between similar cases. The climatic classification chosen for the prototype is from the FAO world soil resources publication [FAO 1993]. The topographical classification is based on FAO's SOTER database classification [FAO 1995] classification. The soil types were classified

according to their texture group. The number of crop types and supporting management practices was limited and therefore they were added to the database exactly as they were described in the literature. It is understood that there will be some variability that cannot be explained by the case-library parameters. However one of the goals of this study is to arrive at a set of features that capture the most important information relevant to impact on water quality. Once determined, the features are used to help the user find other relevant cases. Thus, it is important to have a well-defined template so that cases may be compared. The prediction of possible water quality impacts directly from the selected input parameters is complicated by the dimensionality of the parameters and variation within the input parameter classes and thus is not a goal of this study.

Since the size of the watershed is known to affect the amount of sediment yield at the watershed outlet, it was included as a contributing factor in the database. The watershed is described only by relevant broad-scale features. Therefore, the average or dominant value for the feature is used in the database. The database does not account for any spatial variability that may exist within the watershed. This is not much of a concern for some parameters such as climate where there is little spatial variability when examining average annual values of precipitation and temperature within the watershed. However, other parameters such as soil type may exhibit high levels of spatial variability.

Along with the data used for the case-matching, other pertinent data or metadata is also collected and entered into the database. Although this information is not used for searches, it is available for the user to browse through and obtain more detailed information about the cases in question. Examples of important metadata are when and where the data was collected, and the length of the data set from which average values were derived.

The development of a case library is a time-consuming process. Many cases are required to obtain a good match for a queried case. Some important information may be reported. Many assumptions may be necessary to take a case reported in the literature and make the data "fit" the database model. If a large project were to be undertaken in which studies were performed for such a case matching system, all stakeholders or participants would have to agree on a standard template for the database.

2.2 Tree Induction Algorithms

There exist numerous algorithms for automated decision tree generation [Breiman 1984, Utgoff 1997, Quinlan 1993]. For this study, an incremental variation [Utgoff 1997] of the often cited C4.5 algorithm [Quinlan 1993] was used as the tree generating mechanism. There are three main issues involved with the design of decision trees. There is:

1. The hierarchical ordering of the decision nodes
2. The choice of the partition location
3. Deciding when to finish the tree with a leaf node

One approach to simplify the design of a decision tree induction algorithm is to look at only binary decision trees. In this situation, each decision node corresponds to a single decision on one attribute. The decision space is thus partitioned into hyperplanes orthogonal to the feature axes. This can simplify the tree induction algorithm however, deeper trees are often required.

One technique for locating the positioning of the hyperplanes is based on a measure of the goodness of the partition in terms of a mutual information measure. Consider the partition of a continuous variable or attribute (X) based on some threshold (Φ) and the classification of a set of examples based on (n) classes {C}. The measurement of variable (X) with respect to a threshold can be thought of as a measurement of two possible outcomes (x_1) and (x_2) of event {X}. The average mutual information obtained about the pattern classes from the observation of event (X) can be written as:

$$I(C;X) = \sum_{i=1}^{n}\sum_{j=1}^{n} p(c_i|x_j)\log[p(c_i|x_j)|p(c_i)]$$

Thus the choice of the threshold or test of a variable can be chosen to maximize the average mutual information gain.

Information theoretic based partitioning in this context becomes a matter of searching the attribute space for the partition or test that maximizes the mutual information gain at each iteration of the tree development. These techniques can be used for both continuous input features and discrete features. Many variants of this type of algorithm exist and they often employ a brute force searching strategy.

A number of different stopping criteria have been proposed for eventually generating the leaves of the tree. Some are based on statistical tests, while in the C4.5 algorithm a tree is generated which overfits the data and it is then pruned

2.3 Tree Induction for Case Matching

In our technique, a tree is induced to classify the given cases with respect to discretized sediment yield measurements. Once the tree is constructed the case library can be run through the tree to determine the location of each case within the leaves of the tree. When one wishes to evaluate a potential new case, it can be run through the tree and the cases found in the corresponding leaf can be presented as the matched cases. Examples of this process are included the authors in a longer version of this paper (unpublished), and will be made available on request.

3. CONCLUSIONS

There extensive literature containing watershed studies. However, there are no standard parameters that have been developed to characterize watershed studies that are suitable for large scale indexing of studies. It would be useful to create guidelines and standard reporting parameters that could be included in all studies that would be suitable for this type of indexing procedure.

In the development of complex decision support systems (DSS) which rely on models, the decision-maker using the tool does not usually have the patience (nor the time) to await the outcome of a simulation. a meaningful visualization of that simulation is nevertheless crucial to the confidence in the resulting decision or decisions. Our work has a possible place in the archival and timely retrieval of simulation results in a multi-objective DSS if it is to be more than a presentation or visualization.

4. ACKNOWLEDGEMENT

Support for this research from Youth Horizons Programme of the Government of Canada, Environment Canada, United Nations Food and Agriculture Organization and the Canadian Natural Sciences and Engineering Research Council is gratefully acknowledged.

5. REFERENCES

[Breiman 1984] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group.

[FAO 1993] FAO. (1993). World Soil Resources. An explanatory note on the FAO World Soil Resources Map at the 1:25 000 000 scale. *World*

Soil Resources Reports 66 Rev. 1. Food and Agriculture Organization of the United Nations, Rome

[FAO 1995] FAO. 1995. Global and National Soils and Terrain Digital Databases (SOTER). Procedures Manual. *World Soil Resources Reports 74 Rev.1.* Food and Agriculture Organization of the United Nations.

[Quinlan 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

[Utgoff 1997] Utgoff, P. E., Berkman, N. C., Clouse, J. A. (1997) . *Decision Tree Induction Based on Efficient Tree Restructuring.* Machine Learning Journal. Kluwer Academic Publishers, Boston.