Data Mining in Hydrology

J.M. Spate,[a], B.F.W. Croke[b], A.J. Jakeman[b]

[a]Department of Mathematics, The Australian National University, Canberra ACT 0200, Australia
(jspate@cres.anu.edu.au)

[b]Integrated Catchment Assessment and Management Centre and Centre for Resource and Environmental Studies, The Australian National University, Canberra ACT 0200, Australia

Abstract: A large number of data mining techniques and tools are available for extracting trends, characteristics or rules from data. A selection of those relevant to hydrology are covered in this paper. Clustering, classification, association rule extraction and dominant mode analysis and the ways in which each family of techniques could be used in a hydrological modelling context are considered. In addition to demonstrating the relevance of data mining techniques to the science of hydrological modelling, two specific applications will be discussed. They illustrate the possibilities for improving existing modelling techniques by integrating a data mining approach.

Keywords: *Data; Data Mining; Classification; Clustering*

1. INTRODUCTION

Hydrology can be a data-intensive science. Modelling methods aimed at improving understanding or predictive capacity can require large amounts of observational data in the model building stage, and output similar quantities of data. The observational data used in hydrological modeling and analysis contain measurement and sampling errors, often being collected with imprecise measuring equipment. The quality and nature of collected data are of extreme importance in hydrology, so it follows that all characteristics of such data should be subject to the best possible analysis. Data mining is defined in many different ways in different contexts, but we use this one: *Discovery of interesting, comprehensible and previously unknown rules, trends or characteristics from data.* Essentially, by this we mean discovery of anything that is useful and non-trivial or unexpected from our data.

Typically the minimum data sets used in catchment hydrology are in the form of daily records of rainfall, streamflow, temperature and other climatic variables. Rainfall and streamflow are the two considered here because of their relevance to rainfall-runoff modelling, and also because these quantities do not vary smoothly over long time scales in the way temperatures do. The behaviour of rainfall and streamflow series is more complex and has greater potential to deliver new information.

2. SOME RELEVANT ISSUES IN HYDROLOGY

2.1. Data issues

A number of issues confront us from the collection of such measurements. Firstly, errors arise from the assumption of homogeneous rainfall over a catchment, and the arrival at the total rainfall estimate from some smoothing or interpolation. We will not analyse the interpolation process here, but note that daily rainfall measurement is usually expressed as totals from 9am to 9am. Streamflow records are usually single measurements collected at 12 midnight or they can be averages from one midnight to the next. Hence, the definition of 'daily' varies (for more details see Spate 2002).

Using daily mean streamflow is often a poor representation of the distribution of flows throughout the day. Systematic measurement errors are also a consideration. Rainfall gauges may suffer losses due to evaporation or splashing, and streamflow is intrinsically difficult to measure. Even in a simplified channel like a rectangular weir, where the cross-sectional area of the channel is known, one velocity measurement is not adequate to describe the velocity profile of the flow. In a

natural channel the flow profile is still more variable, and the cross-sectional area difficult to measure.

2.2. Focus of this paper

The data mining techniques discussed here cannot (generally) help remove systematic errors or correct measurement problems. To circumvent this we will change slightly the form of model under consideration. Instead of aiming to predict streamflow volume from total rainfall over the catchment, we will model *what the measured streamflow would be from the measured and interpolated rainfall.* Of course this approach does not remove any errors, but it serves to simplify our analysis.

In addition to providing a quick overview of a few data mining techniques that may be useful to researchers in the modelling community. Examples are chiefly from hydrology, but this is not intended to limit our focus- techniques for extracting additional information from data. We would also like to investigate various methods for dealing with missing data points, a common feature of hydrological data.

3. BASIC TECHNIQUES

A number of techniques are already commonplace in hydrological modelling. Rather than a discussion of existing modelling techniques here, we prefer to suggest new applications of data mining tools. Consider temporal data. There are two ways we can characterise temporal data, as time series or as unordered records of the variables collected on a given day. Using both approaches we can borrow from extensive technical bases established for other applications. For example, some tools used for stock-market analysis can be generalised for other time series, or isolated records can be considered as 'market baskets' and analysed using techniques developed for supermarket purchase analysis.

A third perspective is a middle way between the two above. When examining any hydrograph, we note that the width of most streamflow peaks and some rainfall events have a length of a few days. It seems natural to change the granularity of our time series from days into peaks or events. For the moment we will consider this more granular series as a sequence of shapes, shape A, shape B, and so on. More sophisticated views are discussed later.

One aspect of hydrological data acts in our favour. Daily data, where data mining techniques are perhaps most useful because of the relatively overwhelming quantity of this kind of data and the nature of the information we would like to extract, is usually represented in a sequence of thousands or tens of thousands of records. Most established data mining techniques aim to deal with anything up to millions of records, so the computational efficiency is in general very high for large datasets. In our case, we can take the simplest method without regard for computation time, as it is not likely to be a restrictive factor with our relatively small quantity of data (of the order thousands to tens of thousands of data points). Of course, the small data quantity is not necessarily an advantage. We have less data from which to extract information.

There are a few main data mining concepts we will consider. They are, in approximate order of importance from first to last: Clustering, Classification, Association Rule Finding, Dominant Mode Analysis / Series Similarity Measures. Each will be discussed now in the following sections.

3.1 Clustering

The word 'cluster' has twofold meaning, a cluster being a group of objects, and the verb cluster meaning to group objects according to similarity in some attributes. There are many algorithms available to cluster data, mostly with the object of producing clusters such that the objects within a cluster are as similar, and the defining attributes of each cluster as dissimilar as possible. The number of clusters and level of acceptable dissimilarity in a cluster are considerations in this problem. We require enough clusters so that internal objects are close, but not so many that there are similar clusters. There is a trade-off between the amount of information stored in the clustering regime and understandability or usefulness of that information.

In this section we will examine a basic clustering algorithm, and a few of the problems associated with clustering time series specifically. To illustrate the nature of the clustering problem we will describe in detail one simple clustering method. Called the *k-means* algorithm (Whitten and Frank 1991), we specify beforehand the number of clusters k we want in our regime.

1. Select k different data points at random. These will define our initial clusters.
2. Set the mean value of each cluster to be the value of a chosen point.
3. Assign each data point to the cluster with mean value closest to the data value.
4. Re-calculate the cluster mean taking into account points assigned in step 2.

5. Repeat steps 3 and 4 until the clusters no longer change.

The *k-medoids* algorithm is essentially the same as the k-means, except that where we used means in steps 2-4 we now use medoids. The medoid of a set of point is the n-dimensional analogue of the one-dimensional median. The use of medoids is a sensible choice when the data contains outliers, because an extreme value will not skew the medoid as much as the mean value.

This rough outline of an algorithm could be applied to data in any of the conceptual data forms we described above- isolated records, events or whole time series. However, we must decide how to measure means (or medoids) and distances between shapes, series or records. If we have isolated daily records containing one or two variables, we can just use some Euclidean distance measure. Events and complete series are much more complicated to compare quantitatively.

Consider the set of streamflow peaks:

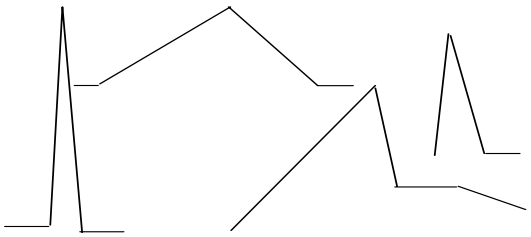

Figure 1. Example Peaks

1. We would like to group peaks according to shape, and we can do this according to the Euclidean distance, but this measure will not reflect similarity between peaks that are the same rough shape but slightly out of phase,
2. Different in magnitude or
3. Start from different base levels.

To obviate these problems, it is sometimes appropriate to normalise each peak, making the mean zero (to remove problem 3) and the variance zero (to remove problem 2). Perhaps the simplest and most effective counter to problem 1 is to choose the starting points of peaks well.

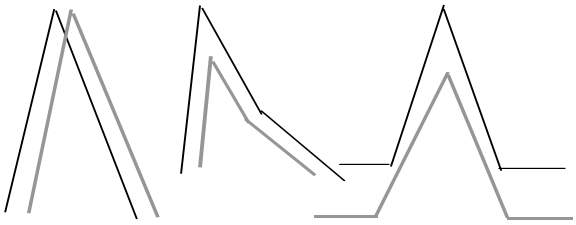

Figure 2. Difficulties in Peak Comparison, Problems 1-3.

Selection of appropriate distance measures between entire time series or sequences of length more than one peak is even more complex. We cannot use Euclidean distance because although the response characteristics and basic shapes may be very similar in two series, the peaks will not occur at the same point in each series (Das *et al*., 1998). There are of course other methods for matching shapes, but those we will examine are a selection of the simplest. Suggested distance measures between series include the distribution of basic shapes, average values, unit hydrographs, and so on. The *unit hydrograph* is a calculated quantity specific to hydrological time series. It is in essence a representative flow response to a pulse of rainfall (see Jakeman *et al*., 1990).

Calculating distance measures between whole series could have wide-reaching applications, and is discussed in a later section (Section 3.4). Clustering of datasets for a large number of catchments can provide regionalisation (extension of model parameters to cover similar catchments) grouping without the need for extensive modelling, and also with less data. In theory, clustering could provide guidelines for replacing small missing sequences. Upon clustering every whole peak, the pieces of information in a peak with holes could be used to find the most likely cluster for the partial peak. The cluster mean or similar representative shape (however defined) could then serve as a template for the reconstruction of the partial peak. This should, on surface analysis, be more accurate than replacing the missing points with information gleaned from the whole set of peaks.

Other potential environmental applications of clustering exist in terrain data analysis (Gallant, personal correspondence). Here identifying and clustering the component shapes in a large-scale Digital Elevation Model (DEM) could lead to the selection of similar geomorphological regions. With sufficient data, the connectivity between geomorphological types would be open for easy examination, and other data that is less simple to

collect such as flow accumulation regions made available. Work is underway in this area.

3.2 Classification

The nature and usefulness of classification is perhaps best illustrated with an example. In a daily rainfall dataset, no information is typically available about rainfall distribution during the day, the rainfall intensity. We do not know if the rainfall measured on a given day fell evenly over the whole 24 hour period or in one intense burst over an hour, say. In the context of daily rainfall-runoff modelling, intensity information would be extremely useful, as the nature of the streamflow responses is dependent on rainfall intensity. The intensity of a rainfall event not only alters the response but is also an important determining factor in the calculation of solute and suspended particle transportation.

High-intensity rainfall data, by which we mean data collected at intervals of order less than a day, is expensive and difficult to collect and hence much rarer than daily data. We would like to find a way of determining the approximate intensity of a rainfall event with daily data. To assist, maximum and minimum temperatures, relative humidities, solar radiation, sunshine hours, evaporation, and other variables may be available on a daily basis.

Using all the daily data series, we can start to hypothesise as to where in the rainfall series intense events occur. Using pre-existing physical knowledge of the catchment or catchment type, it may be possible to find, for example, an approximate set of physical characteristics for a summer storm day with temperature, humidity and rainfall above set thresholds, a date within the likely bounds, and so on. This is clearly a very vague approach that is unlikely to be workable in most areas.

Instead, we select all those catchments in our region of interest where high-intensity rainfall data does exist for at least some temporal interval. Then we can apply some simple criteria to the high-intensity data; for example so much rain must fall in such a small time interval on a given day for that fall to be flagged as an intense event. Having generated a Boolean series with ones on every day with an intense event and zeros elsewhere, we can use a data mining algorithm to automatically extract those combinations of daily data characteristics which tend to occur on a day with a one in the Boolean series.

In the process we will generate a decision tree or other model (see Spate 2002) for automatically deciding whether or not a rainfall event is intense, which can then be applied to daily data where there is not higher resolution data available. The Boolean intensity yes/no marker is called a *classifier* and the process of building and applying the model is known as *classification*. When this process is actually applied to real data useful information can be extracted. In Spate (2002) this is illustrated with the J48 machine learning algorithm in the Weka data mining package (Whitten and Frank, 1991). In the experiments performed, most of the output decision trees were qualitatively logical from a physical perspective, and some branches ending in the 'intense' classifier seemed to be characterising certain physical event types, like summer storms.

The approach behind the J48 algorithm (although the form in Weka is somewhat modified) is a simple divide-and-conquer method (Whitten and Frank, 1991). A decision tree is built up layer by layer by splitting one variable to create two branches. The variable on which the split is performed is located by examining all possible branching schemes, and then choosing the one most likely to give a small tree. This likelihood is measured by the 'purity' of the generated nodes of the tree, or how much information is gained by making the split. The J48 algorithm is discussed more in Spate (2002).

In this way, up to 80% of events defined as 'intense' can be identified in an independent test dataset. Minimal false positives (non-intense events tagged intense) are returned. The building, pruning, and testing of classification models is currently in progress (Spate, in preparation).

There are many other classification algorithms based on entirely different concepts, but the divide-and-conquer is perhaps the simplest. Bayesian methods, neural networks and genetic algorithms are among the conceptual bases of some other approaches. Classification algorithms are currently being trialed for application as land-use models in Thailand (Letcher, personal correspondence). The aim in this project is to discover the rules that govern the choice of crop chosen by farmers under particular economic, environmental, and irrigation conditions.

3.3 Association rule extraction

The concept, although not necessarily the methods, of association rules are similar to classification schemes. In this case we would like to find any rules of the form $A \Rightarrow^{T} B$ that seem to occur in the data with frequency above a given threshold. Here

A and B are just events of a certain type, with the rule *if A occurs then B occurs within time T*. A and B do not necessarily have to be the same variable, for example A could be a rainfall event type and B a shape of flow peak. Here we assume our data is clustered, and in the form of basic shapes. Each basic shape in the series can be replaced by the cluster medoid, mean or representative shape. This yields a sequence of a small number of cluster representative shapes. Each shape in the original sequence is a replaced by a similar shape from a simplified alphabet. Although it represents a less significant saving with our relatively small amount of data, some data compression occurs in this transformation from a series of daily values to a sequence of peak shapes.

Extensions of the basic rule format are possible with a little more computational effort. For example the rule $A_1, A_2, \ldots, A_H \overset{V,}{\Rightarrow}{}^{T} B$ could be interpreted as *if $A_1, A_2, \ldots, A_H$ all occur within time V, then B will occur within time T* (Das *et al*., 1998). The most basic rule form we could consider is $A \Rightarrow B$ with no time dependence, which is sometimes useful.

In our classification example above we had a complete tree with every possible daily record covered at some point by a rule with a classifier. But correspondingly in association rules we obtain only those patterns in time that occur frequently (such as when some small characteristic rainfall event produces similar responses). For example, if a catchment is prone to a short rainfall event with total rainfall of between 10 mm and 30 mm, the stream will usually respond with a particular peak (given normal antecedent conditions) that we will be able to identify. But if a prolonged low-intensity rainfall event occurs over three or four days only a few times in the record, we may not be able to extract a rule that determines the stream response to these conditions.

For a given event type C, say, we may not necessarily have any rule involving C on the left- or right-hand side. Also, where classification in general attempts to build a model that gives the correct classifier for every instance of the training data the rules output from association rule extraction may have exceptions in the training data. What we get from these rules is a partial model composed of common trends.

A partial model may have some application. For example, if we discover from our data that $A \Rightarrow^{1} B$ with 70% confidence (if A, then B occurs in 70% of cases), any missing segments immediately following an occurrence of A can be replaced by B. In this way we may be able to patch some of the holes in our dataset. Reconstruction by rules has advantages over replacement by average values or interpolation. We preserve the prevailing structures in the series.

3.4 Dominant mode analysis and series similarity measures

The technique of extracting dominant modes of a time series is well established. It involves approximation to the series by decomposition with basis functions, usually but not always orthogonal. Fourier analysis is the best known and most used of these. In our example we transform our time series of streamflow into a set of frequencies with amplitudes. The process can be thought of as decomposition into sine waves. Fourier analysis is especially useful in very wet catchments where the flow peaks are not well separated. In these circumstances analysis of individual peaks can be difficult, because of the absence of extractable tail of the response to a single rainfall pulse. Figure 3 shows a few distinct peaks and Figure 4 some overlapping peaks. Fourier mode analyses retain no information about timing of events in the series.

More recently a different, although related, method has emerged. Wavelet transformations spatially decompose the signal into a series of wavelets, the characteristics of which can be carefully tailored. Temporal information is retained, and there is the additional advantage of detailed analysis at varying time scales (Bachman *et al*., 2000). Wavelet transformation has been used to regionalise catchments by taking the wavelet spectra (a form of the results of the wavelet transform) of streamflow records as a signature of catchment response characteristics and clustering on these (Zoppou *et al*. 2002).

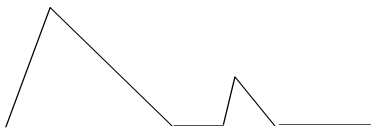

Figure 3. Distinct Peaks

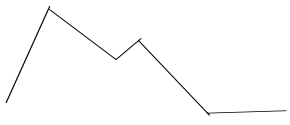

Figure 4. Overlapping Peaks

Other measures can be used to compare the similarity of time series. Usually, the basic

approach (as with Fourier and wavelet analyses) is to map the series onto some low dimensional space and then apply simple distance measures like the Euclidean distance between the mapped vectors (Gunopolis and Das, 2001). Both wavelet and Fourier spectra can be compared this way.

The Longest Common Sequence method calculates the longest continuous subsequence of points that are common between two series. This approach, while enticingly simple, is not directly useful for hydrological data in this form, but when we allow linear scaling and translation of subsequences, it becomes more appropriate. However, hydrological data is not ideal for this kind of analysis - LCS measures are chiefly applicable to data taking discrete values from a small set. Rainfall and streamflow records can be discretised, but will not be considered here.

Some series comparison techniques are probably inappropriate to our problem. Piecewise linear representation is unlikely to be helpful due to the spiky nature of long-term hydrographs, and dynamic time warping, which allows stretching of compression of time axes to achieve better fit between sequences may not preserve information valuable to our analysis. Recently, methods for image database searching have been developed with various mathematically rigorous shape matching bases. For example, the Princeton University Shape Retrieval and Analysis Group image search engine uses the set of distances between a random sample of points inside the shape boundaries to characterise the shape (Chazelle et al., 2002). The approach of the Princeton group can be tested online, with mixed results.

Using distance measures to quantify changes in catchment response characteristics is acceptable practice, but only using simple and/or hydrological modelling concepts like the unit hydrograph or average rainfall/runoff coefficient in the comparison. The development of better distance measures would also be of use in this kind of application.

4. CONCLUSIONS

We have seen a range of hydrological problems that could be phrased as data mining, and methods in common use that also fall under that heading. It is the nature of most hydrological datasets to contain at least rainfall, streamflow and temperature over a period of several years or more, and spatial metadata, making association rule techniques a viable option. The structure of the rainfall and runoff data, decomposable into peaks (although this is not always easy - see Figures 3 and 4) lends itself to clustering and association rule extraction.

Many methods of time series data mining have been developed for stock price analysis, consumer behaviour trend isolation, and even environmental applications. A library of techniques is available for hydrological use, requiring minimal alteration. There are also huge amounts of environmental and hydrological data freely available. These two factors imply that the impact of data mining techniques such as clustering, classification, and association rule extraction could easily have a huge impact on certain hydrological problems.

5. REFERENCES

Bachman, G., Beckenstein, E., and Narici, L., *Fourier and Wavelet analysis.* Springer-Verlag, pp 343-395, 411-475, 2001.

Bloeschl. G. and Sivapalan, M., Scale issues in hydrological modelling: a review. *Hydrological Processes* 9, 251-290, 1995.

Chazelle, B., Dobkin, D., Funkhouser, T., Finkelstein, A., Jacobs, D., Kazhdan, M., Min, P., Chen, J., and Halderman, A., Princeton University Shape Analysis and Retrieval Group. *http://www.cs.princeton.edu/gfx/proj/shape/* (accessed 4/11/2002), 2002.

Das, G., Lin, K.I., Mannila, H., Renganathan, G., and Smythe, P., Rule discovery from time series, *Proceedings of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-1998),* 16-22, 1998.

Jakeman, A.J., Littlewood. I.G. and Whitehead, P.G. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *J. Hydrology*, 117: 275-300, 1990.

Spate, J.M., Honours thesis, Department of Mathematics, The Australian National University, 87 pp., 2002.

Spate, JM. Classification for rainfall intensity extraction. In preparation, May 2003.

Whitten, IH. and Frank, E., *Data mining: practical machine learning tools and techniques with Java implementations,* pp 265-320, Morgan Kaufmann Publishers, 1991.

Zoppou, C., Nielsen, OM. and Zhang, L., Regionalization of daily stream flow in Australia using wavelets and k-means. *http://wwwmaths.anu.edu.au/research.reports/mrr/mrr02.003/abs.html* (accessed 15/10/2002), 2002.