Use of Data Mining Techniques to Identify Crisis in Dryland Living

Saleh A. Wasimi

Faculty of Business and Informatics, Central Queensland University, Rockhampton, Queensland 4702, Australia

Abstract

People living in drylands usually adopt their lifestyle to the prevailing harsh environment around them. Their rudimentary living conditions generally reflect that human lifestyle and environmental factors are intertwined. People in the arid region are used to some variability in their environment as is to be expected from people who live close to nature. However, when this variability extends to the level so as to cause stress to the population, it can be termed 'crisis.' The objective of this study has been to identify the critical values in hydrological, meteorological, environmental, and socio-economic factors that can trigger the onset of crisis.

The study area is the interior drylands of New South Wales in Australia. The region has suffered a number of severe droughts in the last three decades. The amount of data available from direct measurements, remote sensing, secondary sources, and questionnaire survey is phenomenal. Since the volume of data to be dealt with is enormous, and a large part of it is unstructured in the form of textual data, application of data mining tools have been adopted in this study. The standard data mining strategies of classification, clustering, text mining, and association rule mining have been applied.

Essentially, data mining is a black box approach that cannot be conceptualized, but it has found many applications where the problem is too complex or overwhelming with myriad of information. The findings of this study do not unveil any causal relationships as is typical with data mining techniques. Nevertheless, they provide some association of conditions that can forewarn an impending crisis for planning mitigating measures.

INTRODUCTION

The increasing expression of human activity and climate variability require new designs and implementation of integrated instrumentation to address human-climate-terrestrial interactions. Analysis at a regional or catchment scale requires a more holistic vision and far greater commitment than offered by any individual scientific community, and yet for water related issues, should connect to the international policy rulings and recommendations for integrated water resources management, as reflected in, for example, the Water Framework Directive of the European Union (European Commission, 2002). Loucks (2000) identifies that the long-term social well-being in rural communities can only be achieved by aiming for a sustainable use of water resources, which must be an integrating process encompassing climate, natural resources (water, soil), technology, ecology, economy, and society. There are many studies that have been done for integrated modelling of catchments, but they use a systems approach with different modules and combine the results using heuristics, hierarchy, categorization, zoning or abstract principles for a holistic outlook. Krol et al. (2006) provides description of many such modelling examples. Unfortunately, modularization may lead to sub-optimal results (Farnum, 1994). In the absence of any integrated framework that looks at all aspects seamlessly, data mining tools can be used to glean through all information at one time to look for associations and patterns. The focus of this study is to identify the factors which are associated or contributing to the perception of crisis among the population living in drylands using data mining techniques. Crisis is viewed here as the absence of the sense of well-being and is not meant to be extreme distress, misfortune, or victim of a catastrophe.

The paper is organized such that in the next section the perception of well-being is explored. That is followed by a brief description of data mining strategies and tools. After that, the western drylands of New South Wales in Australia is profiled. The paper ends with a discussion.

PERCEPTION OF WELL-BEING

The concept of well-being has been widely researched. Researchers agree that human well-being includes at least: income and satisfaction of basic material needs, the experience of freedom, health and personal security, good social relations, and healthy natural environment (Millennium Ecosystem Assessment, 2003). The concept has a potential to capture subjective understanding of individuals because environment is not ‘given’, but is also created and interpreted by humans. Therefore, the sense of well-being necessarily has to incorporate both subjective and objective facets. Unfortunately, there are too many cases where well-being or quality of life has been modelled in too simplistic a way. A term frequently used to represent quality of life is Human Development Index (HDI), which for example has been presented by Fuhr et al (2003) as a one-dimensional concept of human migration. In an effort for broader meaning, starting with objective indicators available in published literature and through questionnaire surveys of small samples of the rural population of Australia, Table 1 summarises the sense of well-being. The parameters are listed in order of decreased importance.

For the given context, crisis may be considered to exist when the weighted sum of all the factors fall below a certain limit. Each factor may be accounted for using the Likert scale from 1 to 5. It is surprising to see health above everything else. This indicates that people dread sickness above food and water.

Once we identify any given time interval as being either in the state of crisis or well-being, we can add that observation to the data of all other variables – hydrologic, meteorologic, environmental, socio-economic, etc. – for the same time interval and perform data mining of all time series to identify the variables that are significantly associated with the state of crisis.

Table 1. List of significant factors contributing to personal well-being.

Economy and services	Weight	Society	Weight	Environment	Weight
Health Services	0.068	Family health	0.052	Landscape beauty	0.050
Work, Employment opportunity	0.066	Family relations	0.047	Preserving landscape	0.048
Income, farm output	0.060	Personal safety	0.046	Water quality	0.047
Housing	0.049	Community relations	0.040	Air quality	0.045
Recreational facility	0.046	Sports/travel	0.037	Access to nature	0.040
Transport facility	0.043	Family education	0.035	Nature activities	0.040
Educational services	0.041	Political rights	0.026	Hunting, Collecting bush, etc.	0.030
Support services	0.027	Cultural identity	0.017		

DATA MINING TECHNIQUES

Data mining is the process of application of machine intelligence and statistical tools to multifarious data to identify any hidden pattern that may lie in the myriad of data collected on a process or system. The amount of data collected these days from everyday business operations and remote sensing for a region can be phenomenal. Data mining techniques offer the opportunity to scan through huge amount of data to extract any meaningful relationships, patterns or associations. Data mining tools can be applied to both structured, such as temperature records, and unstructured data, such as written texts. Data mining is purely data driven research – it is not empirical research because it does not start with a premise nor it is theoretical research because it does not use postulates or principles.

Data mining can be 'supervised' and 'unsupervised.' Supervised data mining has a dependent or response variable. The two common strategies of supervised learning are classification and estimation. In classification problem, class is the dependent variable and an instance is found to belong to a given class based on its attributes which are the values of the independent variables. In estimation problem, the value of the predictor variable is estimated from the observed values of the independent variables.

Unsupervised learning does not have a dependent or response variable. Two common strategies of unsupervised learning are clustering and association rule mining. In clustering, instances are grouped together based on similarity of their attributes. Clustering can give us an idea as to how many classes belong to a dataset if any. Association rule mining, also called market basket analysis, on the other hand lists objects in groups based on their proximity or associations rather than similarity. It is called market basket analysis because association rule mining endeavours to establish the items that a shopper is likely to buy in one transaction with the purchase of a given item.

There are a number of data mining techniques which are popular with the data mining professionals. The traditional technique for classification is the decision tree. To illustrate decision tree, consider the diagram in Figure 1. The decision to be made is whether to play golf or not. The independent variables are weather outlook, atmospheric humidity and air movement. The path followed from the root node 'Outlook' to a terminal node for set of given values of the independent variables gives the decision at the terminal node. The choice of independent variables and their hierarchy in a decision tree is usually based on entropy concept in information theory. The information content is maximum when the entropy is minimum and the value of the entropy, $H(X)$ of a variable X is given by the following formula:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

where p_i is the probability which can be estimated from frequency of observations.

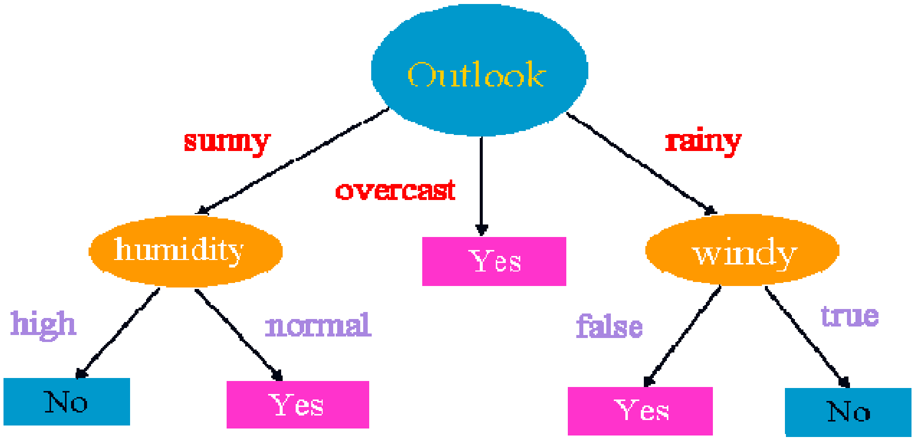

Figure 1. Decision tree for playing golf based on weather outlook.

The most commonly used data mining techniques are artificial neural network (ANN) and support vector machines (SVM). An ANN, which tries to mimic the working of a human brain, consists of a number of interconnected nodes. Figure 2 given the schematic diagram of an ANN, which has been trained for the Exclusive OR (XOR) function. The training process is to find the value of the path weights so that training data can be reproduced. Table 2 gives the training data for the XOR function.

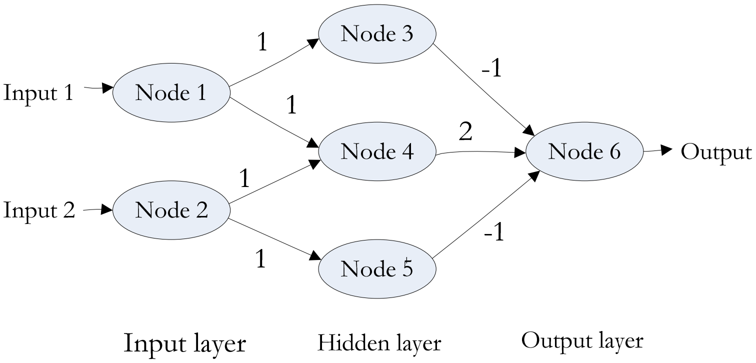

Figure 2. An artificial neural network trained for the XOR function.

Table 2. Boolean XOR problem.

Input 1	*Input 2*	**XOR**
0	0	0
0	1	1
1	0	1
1	1	0

Each input is entered into a node in the input layer. It then transmits it to nodes in the next layer multiplied by the path weight. The node in the next layer, which is usually a hidden layer, aggregates all the incoming inputs passes it to the next layer either unchanged or transformed by a sigmoid or threshold function. In Figure 2, the hidden nodes use a threshold function: output = 1 if sum of inputs is greater than or equal to 1, otherwise 0. The path weight gets multiplied to the output from a node before that is transmitted to the next set of nodes. The output node usually aggregates the inputs and gives the result untransformed, but it can also use a threshold function. XOR is a difficult relationship to write in an equation form, but an ANN can easily reproduce it.

Another popular data mining technique is support vector machine. It is mathematically more involved that ANN. To illustrate the concept Figure 3 is presented where two classes represented by round and square dots exist. There are many lines that can be drawn to separate the classes. The middle separating line, called the hyperplane, can be flanked by two parallel lines on either side representing the margins until each of the margin lines touches a point. When the width of the margin is maximum as shown in Figure 3 (b), the hyperplane is known as the optimal hyperplane. The instances or cases which touch the margins of the optimal hyperplane are called the support vectors. SVM only works with the support vectors for analysis, thereby reducing drastically the computational requirements.

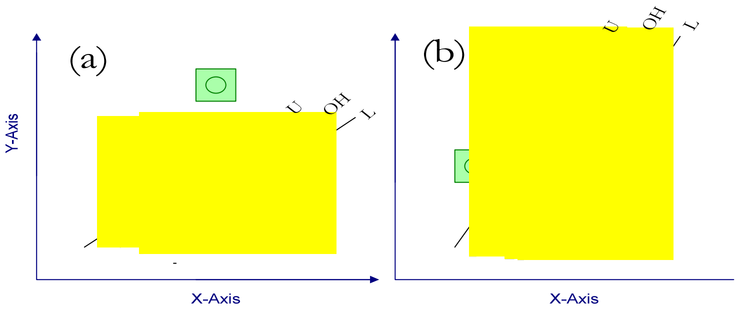

Figure 3. Instances from two different classes. Margins are narrow in (a), but maximum in (b).

For estimation problems, regression analysis is quite commonly used. Since data mining potentially deals with a lot of data and numerous variables, it is required to do some pre-selection to avoid computational problems such as multicollinearity problem.

DROUGHT IN NEW SOUTH WALES

Figure 4 depicts the drought declared areas in New South Wales in 2006. The magnitude of drought conditions defies imagination. The data collection of the drought conditions is still an ongoing process. After the data collection phase is complete attempts will be made to correlate the various factors and try to identify the best strategies to deal with the drought.

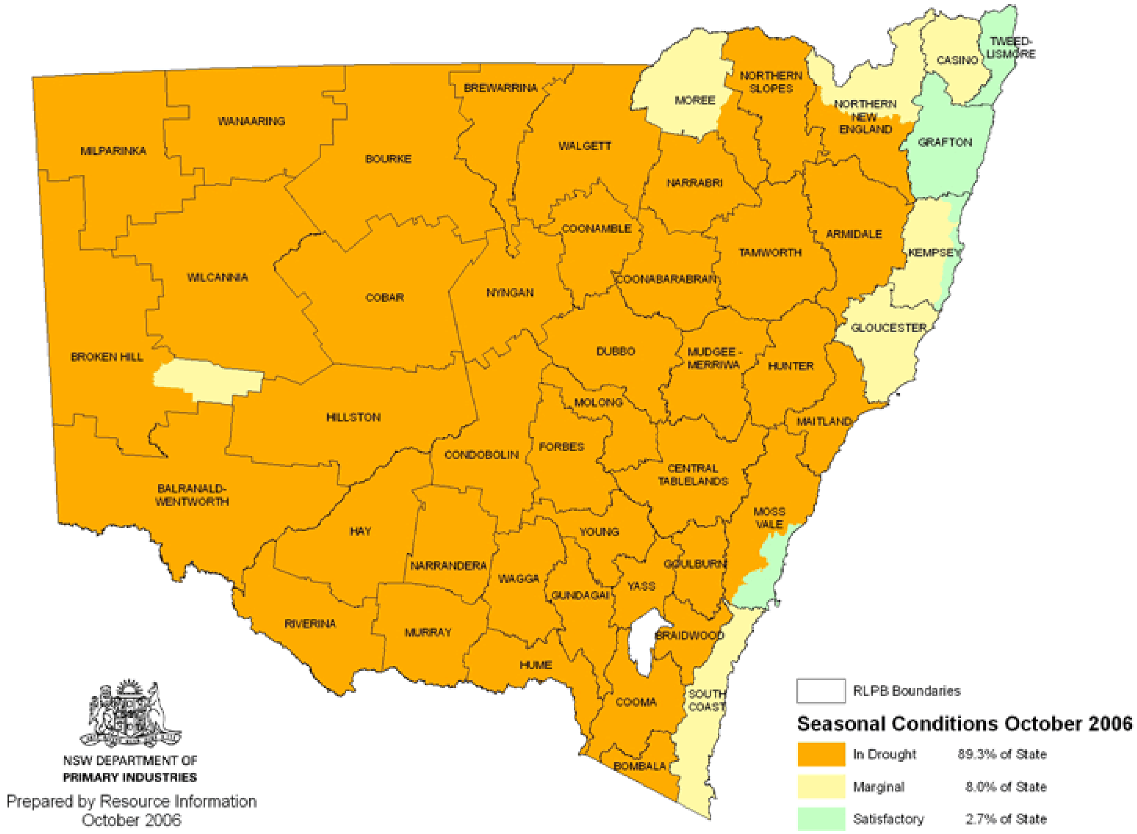

Figure 4. Map showing the severe extent of drought in New South Wales.

At present drought is defined as, "Events that occur once in every 20 to 25 years and has an impact for more than 12 months." Once are area is declared to be in drought government assistance is available in the form of relief payment and interest rate subsidies so that farmers are not forced to leave their land.

It is identified that dealing with drought is very much dependent on the limiting farm resources, which primarily comprise:

Mental and physical energy to do the tasks.

Funds available.

Stock and domestic water available.

Feed reserves (paddock and stored) available.

Surface/subsoil moisture available for crop growth.

Livestock fat reserves stored enabling controlled weight loss and

Need to service farm machinery.

One of the major challenges of New South Wales farmers is to deal with their farm animals. Depending on the scale of the drought and the availability of resources the following advice are in place by the Department of Natural Resources for the affected farming communities.

Selling stock – Disposal of some stock early in drought brings in good price while the animals are still healthy.

Production feeding – Keep a healthy breeding stock and grow animals when they would sell for a premium price.

Maintenance feeding – animals allowed to use some of their body reserves before commencing maintenance feeding.

Agistment – Transferring animals to a suitable region.

Trading in livestock – this may result in loss of genetic base and risk of introducing disease.

Humane destruction – It is against law to starve livestock to death.

CONCLUSIONS

The objective of this study has been to establish relationship between various factors for the drought conditions in New South Wales. The data collection process is ongoing. Data exists in various formats such as textual data, satellite imagery, numeric data and anecdotal data. Data mining techniques can seamlessly process all this information to get an integrative look of the drought conditions.

REFERENCES

European Commission (2002), "Tap into it! The European water framework directive." Bruxelles.

Farnum, N.R. (1994), *Statistical Quality Control and Improvement*, Duxbury Press, Belmont, California, pp. 500.

Fuhr, D., Grebe, M., Doring, A, da Rocha, F.M., and Lantermann, E. (2003), "Quality of life and migration concepts and results of the socio-economic survey in Tauá and Picos." In: T. Gaiser, M.S. Krol, H. Frischkorn and J.C. de Araújo, Editors, *Global Change and Regional Impacts. Water Availability and Vulnerability of Ecosystems and Society in the Semi-Arid Northeast of Brazil*, Springer-Verlag, New York (2003), pp. 349–360.

Krol, M., Jaeger, A., Bronstert, A. and Guntner, A. (2006), "Integrated modeling of climate, water, soil, agricultural and socio-economic processes: A general introduction of the methodology and some exemplary results from the semi-arid north-east of Brazil." Journal of Hydrology. Vol. 328, Issue 3-4, September, pp. 417-431.

Loucks, D.P. (2000), "Sustainable water resources management." Water International. Vol. 24, No. 1, pp.3-10.