Author(s): Hafeez, M.M. ; Islam, M. ; Khan, M.A.

Title: Irrigation Water Demand Forecasting - A Data Pre-processing and Data Mining Approach Based on Spatiotemporal Data

Conference: Australasian Data Mining Conference (AusDM)

Location: Ballarat, Australia **Date:** 1-2 December, 2011

Year: 2011 **Pages:** 183 - 194

Editor: Vamplew, P., Stranieri, A., Ong, K.-L., Christen, P. and Kennedy, P. J.

Publisher: Australian Computer Society

Place of Publication: Sydney, Australia

Abstract: World population is increasing at a fast rate resulting in huge pressure on limited water resources. Just about 3% of the earth¿s total water is freshwater that can be used for various applications including irrigation. Therefore, an efficient irrigation water management is crucial for the survival of human being. In our study area farmers need to order water based on their requirements. Once a request for water is made it typically takes about 7 days to get it at the farm gate from the upstream ...

Irrigation Water Demand Forecasting – A Data Pre-Processing and Data Mining Approach based on Spatio-Temporal Data.

Mahmood A. Khan [1], Md. Zahidul Islam [2, 3], Mohsin Hafeez [1]

[1] International Centre of Water for Food Security, Charles Sturt University, Wagga Wagga 2678, NSW, Australia
[2] School of Computing and Mathematics, Charles Sturt University, Wagga Wagga 2678, NSW, Australia
[3] Centre for Research in Complex Systems (CRiCS), Charles Sturt University, Bathurst 2795, NSW, Australia

makhan@csu.edu.au, zislam@csu.edu.au, mhafeez@csu.edu.au

Abstract

World population is increasing at a fast rate resulting in huge pressure on limited water resources. Just about 3% of the earth's total water is freshwater that can be used for various applications including irrigation. Therefore, an efficient irrigation water management is crucial for the survival of human being. In our study area farmers need to order water based on their requirements. Once a request for water is made it typically takes about 7 days to get it at the farm gate from the upstream. Therefore, farmers need to estimate water requirement for the next 7 days in advance in order to get it at the farm gate on time. Currently there is no reliable tool available to the farmers of our study area for estimating future water requirement accurately. Hence, a water demand forecasting technique is crucial for the efficient use of available water.

In this study we first prepare a data set containing information on suitable attributes obtained from three different sources namely meteorological data, remote sensing images and water delivery statements. In order to make the prepared data set useful for demand forecasting and pattern extraction we pre-process the data set using a novel approach based on a combination of irrigation and data mining knowledge. We then apply a decision tree technique to forecast future water requirement. We also develop a web based decision support system for the managers, farmers and researchers in order to access various data including the prediction of possible water requirement in future. We evaluate our pre-processing technique by comparing it with another approach. We also compare our decision tree based prediction technique with a traditional prediction approach. Our experimental results indicate the usefulness of our pre-processing and prediction techniques.

Keywords: Demand forecasting, Data Mining, Decision Tree, Decision Support System, Water management, and Data pre-processing.

1 Introduction

Water availability plays an important role in agricultural. The world population is growing at a fast rate resulting in rising demand for household and irrigation water. Therefore, in the past decades irrigation water supply systems are under huge pressure in fulfilling the irrigation water requirements. Over 70% of the water in Australia is currently being used by agriculture (Khan et al. 2009). Since all the existing water resources are fully exploited and it is not possible to extract more water, the best alternative is to increase the water productivity.

For efficient irrigation water management, application of various hydrological models and data mining approaches has become crucial. Most of the water delivered for irrigation is not always efficiently used for crop production. On an average only 45% of the water is used by crop, 15% is lost during conveyance, 15% is lost in supply channels within the farms and the remaining 25% is lost due to inefficient water management practices (FAO 1994, Smith 2000). Therefore, it is evident that most of the water losses occur at farm level because of inefficient water management practices. In order to increase the water management efficiency a water demand forecast model can be useful.

There is a propagation delay for water to reach a farm from the original source in the upstream. Often the delay can be as long as 7 days, as it is the case for many farms in our study area. Therefore, to get water on time a farmer often needs to order water 7 days in advance. Since currently there is no reliable scientific tool for the farmers at our study area for estimating exact water requirement, a farmer relies on his/her experience for guessing the possible water requirement for the next 7 days. Hence, a farmer generally either overestimates or underestimates the water requirement. If the requirement is overestimated there will be on farm water loss, whereas if it is underestimated there can be adverse effect on the crop productivity. Therefore, having a reliable water demand forecast model can be useful for a farmer to estimate water requirement more accurately. The demand forecasting tool can also be useful for the irrigation managers for estimating water requirement for the whole irrigation area.

There are two major approaches for estimating water demand: i) conceptual and ii) system theoretical (Pulido-Calvo et al. 2009, Zhou 2002, Alvisi 2007). A conceptual model predicts the irrigation water requirement based on several factors including soil moisture, seepage, and evapotranspiration. Subsequently irrigation managers use these factors to estimate irrigation water demand for the whole season. However, water requirements estimated at the beginning of the irrigation season may not be the same as the actual water usage due to many reasons such as difference in expected and actual weather conditions and change in farming practices (Pulido-Calvo 2003).

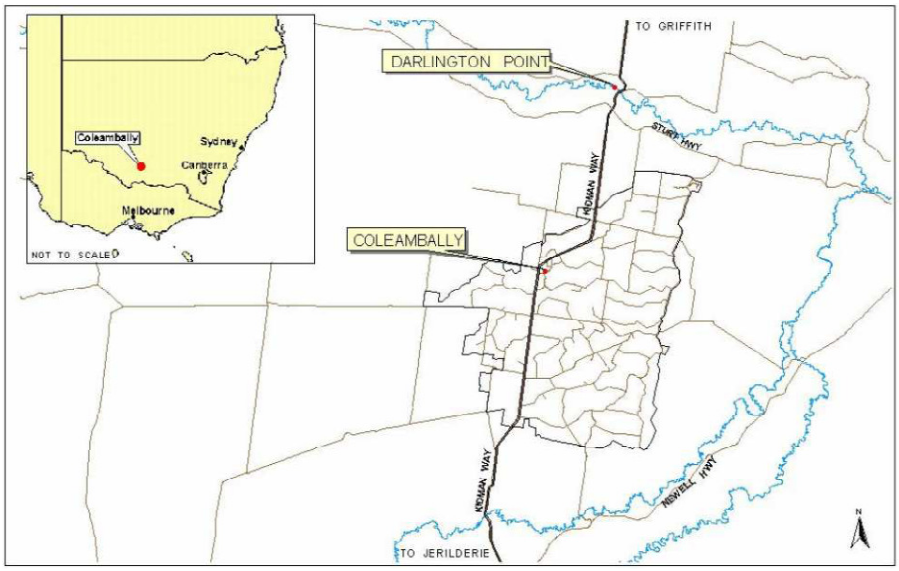

Figure 1: Location of Coleambally Irrigation Area (Source: CICL Annual Compliance Report, 2010)

The second approach for estimating water demand is known as system theoretical approach. In this approach a model is first trained on available data and then used for estimating future water demand. The system theoretical approach is more efficient and accurate than the conceptual approach (Pulido-Calvo 2008). Moreover, it can base on easily available data only.

To build an effective model using data mining techniques adequate historical data for the parameters such as crop water usage, crop type and weather conditions are required. A decision tree model can be useful for water demand forecasting. A decision tree (as shown in Figure 5) contains nodes and leaves, where each node in the tree tests an attribute and each leaf represents the value for the records belonging to the leaf (Han & Kamber 2001).

It recognises the relationship between the classifying (class) and the classifier (non-class) attributes. A class attribute is an attribute within the data set, which contains the values that are possible outcomes of the record. A decision tree analyses a set of records whose class values are known (Quinlan 1996). In other words, a decision tree explores patterns also known as logic rules from any data set (Islam 2010).

In this paper, we perform few interesting data pre-processing approaches using a combination of knowledge on irrigation engineering and data mining in order to improve the quality of our training data set. During the data pre-processing step we compare a few different approaches and finally adapt the most logical one. Our training data set contains attributes on various weather parameters (such as maximum and minimum temperature, wind speed, humidity, rainfall, and solar radiation), soil type, crop type (obtained from remote sensing satellite images) and water usage.

We then build a decision tree on the pre-processed training data set in order to extract existing pattern and predict future water demand for the next 7 days. The performance of the decision tree model is compared with a traditional way of estimating water demand using actual evapotranspiration (ET_c).

Our experiments indicate that as a result of the data pre-processing the quality of the training data set increases significantly. Moreover, the accuracy of water demand prediction made by the decision tree approach from the pre-processed data is higher than the accuracy of the traditional approach.

The demand forecasting technique is incorporated into our DSS called Coleambally IRIS (Integrated River Information System). Coleambally IRIS is a web based information system which stores information about time series, geospatial, climate and remote sensing data. It helps the users in decision making for sustainable irrigation water management. The paper is organised as follows: In section 2 we introduce our study area. Data collection, data pre-processing and the decision tree based demand forecasting technique are explained in section 3. Section 4 introduces our web based Decision Support System for managers, farmers and researchers. Experimental results are discussed in section 5. Finally Section 6 provides concluding remarks.

2 Study Area

The Coleambally Irrigation Area (CIA) is situated in the Riverina District of New South Wales which falls under Murrumbidgee River catchment as shown in Figure 1. CIA was developed in 1970 when the Water Conservation and Irrigation Commission acquired a large number of pastoral holdings to make use of water diverted from the Snowy Mountains Hydro-Electric Scheme. CIA contains approximately 79,000 hector of irrigated agriculture.

Coleambally Irrigation Cooperative Limited (CICL) was formed in the year 2000 after privatisation of irrigation corporations. Water in CIA is used to irrigate crops on 473 irrigation farms with an average size of 250ha. The main summer crops grown in CIA from November – April include rice, soybeans, maize (corn), grapes, prunes, sunflowers and Lucerne, whereas the winter crops grown from May – October include wheat, oats, barley, canola and Lucerne. Pasture for grazing is generally grown round the year.

Different soil types found in CIA are highly suitable for irrigated cropping, for example heavy clay is suitable for production of rice. Soil types such as, Self Mulching Clay (SMC) with a small portion of Sand are found at the northern areas of CIA, while Red Brown Earth (RBE) and Traditional Red Brown Earth (TRBE) are generally found in south as shown in Figure 3.

The climate of the area ranges from warm temperature with hot summers and mild winters. The climate averages obtained over the last 10 years from the Bureau of Meteorology (BoM) weather station located at Coleambally showed an average maximum temperature in January of 33.2°C and an average minimum temperature in July of 3.6°C. The long term average rainfall is 396 mm per year.

The surface water distribution for CIA from Murrumbidgee River is through the Main Canal and supply channels of length 477kms. Due to drought in the last decade, there has been a significant reduction in water allocations to the farmers highlighting the need to manage water demand and supply more sustainably in CIA.

According to the Annual Compliance Report of CICL (Coleambally Irrigation Company Limited, 2010), the average water allocation in 1995-2001 was 91% and was significantly declined to only 15% in 2006-2009. Due to declining water allocation and changing weather patterns, CIA requires new management measures for water use efficiency ranging from the farm (sub system) to the whole irrigation area (system) level. These measures can be enhanced by developing a model for irrigation water demand forecast which helps CIA farmers and managers use available water efficiently by irrigating right amount of water at the right time.

2.1 Existing water management practices in CIA

CICL is owned and operated by local farmers. It runs on a cooperative status. Its main purpose is to manage surface water distributions to the farms in CIA (Jackson 2009). CICL supplies water to all the farms through supply channels using a demand driven irrigation system. The water supply is managed every year based on the following rules.

1. Every farm has a predefined water entitlement which remains the same over the years. Based on the water availability in a particular year and the entitlement of a farm the allocation of water for the farm is determined.
2. Water is delivered by CICL to farmers upon their request.
3. Water ordering is made by the farmers based on their knowledge and the experience.

The CICL officers use their domain knowledge, experience and water order information placed by farmers to order water from the state water commission which supplies irrigation water to CIA. It is a common practice among the CIA farmers to place the orders with their initial plans on the crop type and the irrigation area at the beginning of a season. Often, the farmers revise their earlier decision (change their mind) after placing the orders, depending on water availability, market prices and many other socio economic factors. Currently in CIA there is no particular method or model to estimate the future water requirement, except the method using Evapotranspiration (ET). This method is generally used by the irrigation managers to estimate water demand. The water ordered by the farmers does not provide CICL with adequate information about the actual cropping area, which may result in inaccurate water ordering by the managers from the state water. It is also not mandatory for the farmers to provide information regarding the area under different crops. Due to the inaccurate estimate of water demand, shortage of water on the day of delivery can occur. An irrigation water demand forecast model can help overcome these difficulties by using remote sensing and meteorological data to ensure the optimum quantity and timely delivery of water for crops.

2.2 Irrigated Crop Area and Land Use in CIA

In 2009/10, rice and corn were the major summer crops covering 11.2% of the total irrigated area, while wheat, barley, canola and pasture being the main winter crops covering 65.8% of the total irrigated area. These crops have remained dominant since the establishment of CIA (CICL ACR, 2010). Table 1 illustrates the information about the crop area and the proportion of water used by them in 2009/10. Figure 2 shows the land use and land cover map obtained by remote sensing for the same year.

Crop	Area(ha)	Percentage of water delivered by CICL (%)
Rice	3668.5	46
Corn/Maize	1516	4
Wheat	10635	10
Barley	10499	7
Canola	2523	2
Pasture	6903	12
Others *	10995	19
Total	**46431**	**100**

Table 1: Crop area and percentage of crop water use for the year 2009/10 (Source: CICL Annual Compliance Report 2010)

* Lucerne, Triticale, soybean, sunflower, clover, prunes, grapes, vegetables etc.

3 Data Collection and Data Pre-processing

To build the training dataset, we collect data from three different sources. The first source is the water delivery statements that are obtained from CICL and provides us with the information about total water usage for a crop growing season by each farm. The second source is the meteorological data that are obtained from the installed weather stations in the study area. The third source is spatial data that are of two types a) Land Use Land Cover images, which provide us with the information about the crops grown and the cropping area as shown in Figure 2, b) soil type images that gives us the information about the different soil types associated with the farms in the study area as shown in Figure 3.

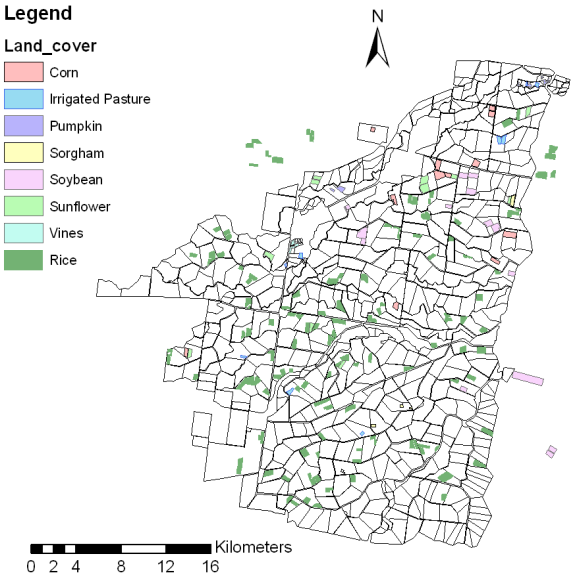

Figure 2: Land Use and Land Cover map of CIA for summer 2009/10

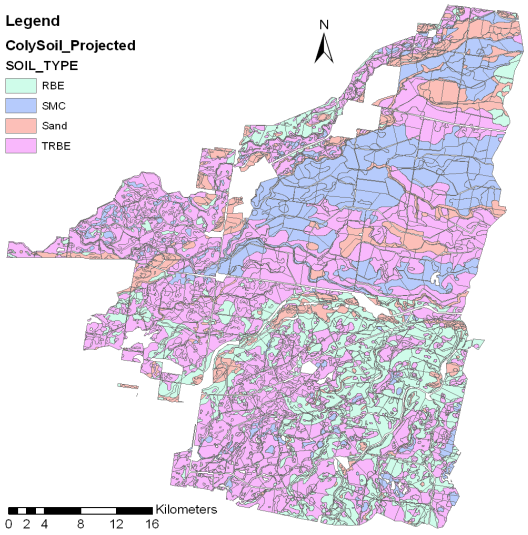

Figure 3: Different soil types of CIA

To run the decision tree algorithm, there is a strong need for data pre-processing to prepare good quality data. Data pre-processing takes approximately 80% of the total data mining effort (Zhang et al, 2003). It is also known that good results can be achieved by using data mining techniques/algorithms only if we have a good quality data set (Miksovsky et al, 2002). Often the real time data is very inconsistent as it contains many attributes which are not useful for our purpose and have some missing values. The purpose of data pre-processing is to remove noise from the data, extract and combine the required/relevant attributes from different data sources, make the data reliable and transform the data into our required format (Xu, 2003). By pre-processing the raw data, it is possible to prepare a good quality data set, which enables efficient and quality knowledge discovery (Zhang et al, 2003). Using the raw data obtained from three different sources we prepare our training data set as follows.

3.1 Attribute Selection

Based on our domain knowledge in irrigation, we select the attributes that have major influence on crop water usage. Moreover, we only select those influential attributes the data for which are easily available throughout the crop growing season. The selected attributes are various weather parameters (such as Maximum and Minimum Temperature, Wind speed, Humidity, Rainfall, and Solar Radiation), soil type, crop type and crop water usage.

3.2 Construction of data set

The data set was prepared from the historical data obtained from three different sources as discussed before. In this study we consider the data set as a two dimensional table where columns are attributes (categorical and numerical) and rows are records. Each record holds the daily average values of the corresponding attributes. Categorical attributes include soil type and crop type, whereas all other non-class attributes are numerical.

The water delivery statement only provides us with the information on the date and quantity of water supplied to a farm. Note that a farm does not take water supply every day. Instead it takes a specific volume of water on a day and uses the water for a period of time. Generally the farms have their own water storage facilities. The farmers can then order more water when they require. Therefore, from the water delivery statement it is not possible to estimate the exact amount of water usage for a particular day. However, our training data set contains records having daily average values of the non-class attributes. Each record represents information on a farm and a farm can have many records in the training data set. Hence, in order to obtain an accurate relationship between the non-class attributes and the class attribute (i.e. water usage) we need to store daily water usage for each record of the training data set.

We consider three possible techniques/approaches to estimate the daily water usage of a farm. We call the techniques as Equal Water Distribution, Averaging Out of the Parameters, and Reference Evapotranspiration Based Estimate. We explain the techniques as follows.

In equal water distribution technique we divide the volume of water delivered to a farm by the number of days between two consecutive deliveries. Therefore, we get an average water usage per day. However, if we divide the water usage evenly among the days then water usage remains same for each day regardless of weather conditions. Since crop water demand depends on the climatic conditions this approach does not appear to be a suitable one for this study.

In the second approach we take an average of both weather parameters and water usage for the days between the deliveries. Thus we convert the records representing the days between the deliveries into a single record having average values as shown in Figure 4. For example, let us assume that 100Mega Litre (ML) of water is delivered to a farm on the 1st of October and 400 ML of water is delivered on the 15th of October. In this approach we take the average weather parameters of the 15 records

and convert them into one record as shown in Figure 4. In this case the water usage of the record is 100/15 ML.

This approach appears to be a little better than the first one since we take average of the weather parameters along with the water usage. Therefore, water usage is not same among all the days having very different weather conditions. However, we introduce noise and also lose information due to the averaging out of the values. This is similar to the problem we usually face due to generalisation of data.

For example, let us consider two records R1 and R2. For R1 T-Max is 18^0 C and Humidity is 20%. For R2 T-Max is 38^0 C and Humidity is 80%. Let us also assume that the total water supply for the two days is 0.16 ML/ha. According to the first approach total water usage will be equally divided among the records. Both R1 and R2 will have 0.08 ML/ha water usage even though they have significantly different weather conditions. The second approach will merge the records into a new record say R3 having average values. R3 will have T-Max as 23^0 C, Humidity as 50% and Water Usage as 0.08 ML/ha. If the original water usage for R1 is 0.01 ML/ha and R2 is 0.15 ML/ha then both approaches appear to be unsuitable for extracting relationships between weather parameters and water usage. However, it is clear that the second approach preserves the relationship a little better than the first one.

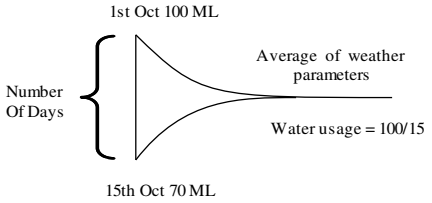

Figure 4: Averaging out of the parameters approach

Moreover, in the above example we assume that the water that was delivered previously has been used to irrigate the crops before the next delivery of water. However, in reality a farmer can actually order water while still having some water stored from the previous delivery. Similarly a farmer can also delay the water ordering. Therefore, dividing the whole amount of water that was delivered last time by the number of days between two deliveries may not get the right amount of water usage per day.

We come up with the third approach to resolve the issues with the first two approaches. In this approach we take Reference Evapotranspiration (ET_o) factor into consideration. Crop water usage can be calculated through Evapotranspiration (ET) which is the product of crop coefficient k_c and reference evapotranspiration (ET_o) (Al-Kaisi and Broner, 2009). Each crop has a constant crop coefficient value for a specific growth stage.

Let n be the number of days between two water deliveries for a farm, W_T be the amount of water delivered for the n days, and W_i be the water usage for the ith day. Note that W_T is the amount of water delivered at the beginning of the n days, and not the summation of two deliveries for n days. We obtain the daily ET_o values, for all n days, from our weather stations in the study area. We then calculate the coefficient x_i for the ith day, where $x_i = \frac{ET_o^i}{\sum_{j=1}^{n} ET_o^j}$. ET_o^i is the ET_o of the ith day. Finally W_i is calculate by multiplying x_i and W_T, i.e. $W_i = x_i \times W_T$.

There are several advantages of the third approach. Unlike the first approach here we do not use the same average water usage for the days having different weather conditions. Moreover, unlike the second approach it does not average out the weather parameters and water usage for the days in order to generalise the records into one. It estimates water usage, as accurate as possible, for each day and thereby uses each record of the training data set.

The final part of our data set is prepared by gathering the information from the spatial images for seasonal land use (cropping pattern) and to determine soil type of the farms. By using the spatial maps processed from satellite images we extract the crop type, cropping area and soil type information of every farm for a particular season.

In order to reduce the inconsistency of our data set we neglect the data from the farms having more than one soil type and consider only those farms with homogeneous soil type. Each record in our final data set holds daily average values of the weather parameters and crop water usage. However, values for crop type and soil type remain the same for whole growing season. Our data pre-processing technique uses a combination of the knowledge on data mining and irrigation engineering.

3.3 Application of a Decision Tree Algorithm on our Pre-processed Training Data Set on CIA

A decision tree algorithm is applied on the pre-processed data to extract the relationship between the non-class attributes and crop water usage. To generate a decision tree from our data set we consider crop water usage as the class attribute and all others as non-class attributes as shown in Table 2. Crop water usage is considered as a categorical attribute. While generating the decision tree, when an attribute is tested for a node if the attribute is numerical then there are two branches for the node (i.e. the data set is divided into two mutually exclusive horizontal parts). One branch contains all the records ">k" and the other contains all the records "<=k" of the data set, where k is a constant and it is one of the values of the attribute. However, if the attribute tested is categorical then there are n branches for the node, where n is the number of distinct values of that attribute.

We implement C4.5 algorithm to generate a decision tree on our pre-processed training data set. C4.5 algorithm takes a divide and conquers approach to build a decision tree from a training data set using information gain (Quinlan 1993).

We briefly introduce C4.5 algorithm as follows (Quinlan 1993, Islam 2010).

D – Whole data set. A two dimensional table where columns are attributes, and rows are records. Each record contains related information about the attributes.

T- An attribute with n number of mutually exclusive outcomes T_1, T_2……. T_n.

c - Number of classes i.e. domain size of the class attributes C.

p (D, j) - proportion of records in D belonging to the j^{th} class.

$D_i \subseteq D$ - the horizontal partition of the data set where all the records have T_i for the attribute T.

p (D_i, j) - proportion of records in D_i belonging to the j^{th} class.

|D| - size of the data set D.

$|D_i|$ - size of the partition D_i.

Step1: Entropy Calculation

The algorithm first calculates the entropy (a measure of the uncertainty associated with the class values of a set of records) of the whole dataset D using the following equation.

$$I(D) = -\sum_{j=1}^{c} p(D,j)\log_2\big(p(D,j)\big) \quad (1)$$

Gain Ratio Calculation for a Categorical Attribute T:

Step 2A: The algorithm then calculates entropy for a subset of the data set where all records have T_i for T as follows.

$$I(D_i) = -\sum_{j=1}^{c} p(D_i,j)\log_2\big(p(D_i,j)\big) \quad (2)$$

Step 3A: Weighted Entropy of the whole data set when attribute T is tested

$$I(D,T) = \sum_{j=1}^{n} \frac{|Di|}{|D|} \text{ X } I(Di) \quad (3)$$

Step 4A: Gain of an attribute T can be calculated by subtracting the weighted entropy from the total entropy of the dataset.

$$Gain(D,T) = I(D) - I(D,T) \quad (4)$$

Step 5A: Split Info of each attributed is calculated as follows.

$$SplitInfo(D,T) = -\sum_{j=1}^{n} \frac{|Di|}{|D|} \text{ X } \log_2\left(\frac{|Di|}{|D|}\right) \quad (5)$$

Step 6: Gain Ratio

$$GainRatio(D,T) = \frac{Gain(D,T)}{SplitInfo(D,T)} \quad (6)$$

Gain Ratio Calculation for a Numerical Attribute T:

When T is a numerical non-class attribute, the records are first rearranged so that all values of T are placed in ascending or descending order. Now the data set is horizontally divided into two parts, D_1 and D_2, based on a splitting point value *k* so that the domain of T in D_1 is [*l*, *k*], where *l* is the lowest value of the domain, and D_2 is [*k*+1, *u*], where *k*+1 is the next higher value to *k* and *u* is the upper value of the domain.

Step 2B: I(D,T) is calculated as follows.

$$I(D_i) = -\sum_{j=1}^{c} p(D_i,j)\log_2\big(p(D_i,j)\big), \text{for} 1 \le i \le 2 \quad (7)$$

$$I(D,T) = \sum_{j=1}^{2} \frac{|D_i|}{|D|} \text{ X } I(D_i) \quad (8)$$

The splitting point for which we achieve the minimum I(D,T) is considered as the best splitting point for T.

Step 3B: Gain can be calculated by subtracting the weighted entropy from the total entropy of the dataset.

$$Gain(D,T) = I(D) - I(D,T) \quad (9)$$

Step 4B: Split Info

$$SplitInfo(D,T) = -\sum_{j=1}^{2} \frac{|D_i|}{|D|} \text{ X } \log_2\left(\frac{|D_i|}{|D|}\right) \quad (10)$$

Step 5B: Gain ratio is calculated as follows.

$$GainRatio(D,T) = \frac{Gain(D,T) - (\log_2(N-1)/|D|)}{SplitInfo(D,T)} \quad (11)$$

where, N is number of distinct values of attribute T. Figure 5 shows a part of a decision tree obtained from our training data set.

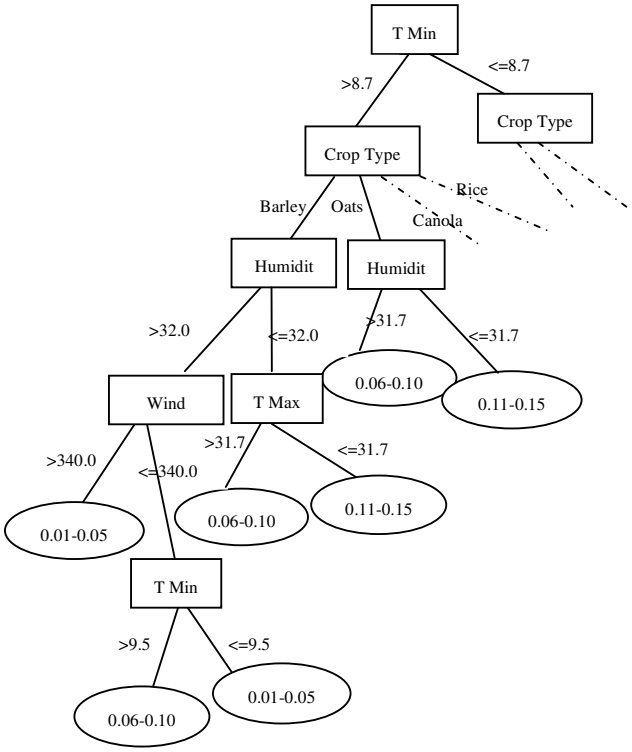

Figure 5 : Part of Decision Tree generated by C4.5 algorithm on our training data set

4 Demand Forecast Technique Implemented in our Web based Decision Support System (DSS)

The decision tree algorithm for irrigation water demand forecasting is developed in java. The demand forecasting model is incorporated in Coleambally IRIS DSS. Within the DSS, PHP (hypertext pre-processor) and java interacts with each other to generate a Decision tree and predict future water usage values from the decision tree. PHP calls the relevant java files which automatically generate a decision tree in order to perform a future prediction. The decision tree and the water demand prediction is displayed in the web pages. Description of the interaction between PHP and java is demonstrated by the conceptual diagram as shown in Figure 6.

Tmax (°C)	Tmin (°C)	Humidity (%)	Wind Speed km/day	Rainfall (mm)	Solar Radiation (MJ/m^2)	Soil Type	Crop Type	Crop Water Usage (ML/Ha/day)
18.1	3.8	80	122	0.2	9.5	SMC	Barley	0.01-0.05
16.4	6.7	48	481	0	16.6	RBE	Wheat	0.06-0.10
30.1	14.0	65	275	0.0	24.7	SMC	Rice	0.11-0.15
30.7	15.9	58	257	0.0	29.3	SMC	Corn	0.06-0.10

Table 2: Example of our training data set, Crop Water Usage is the Class attribute.

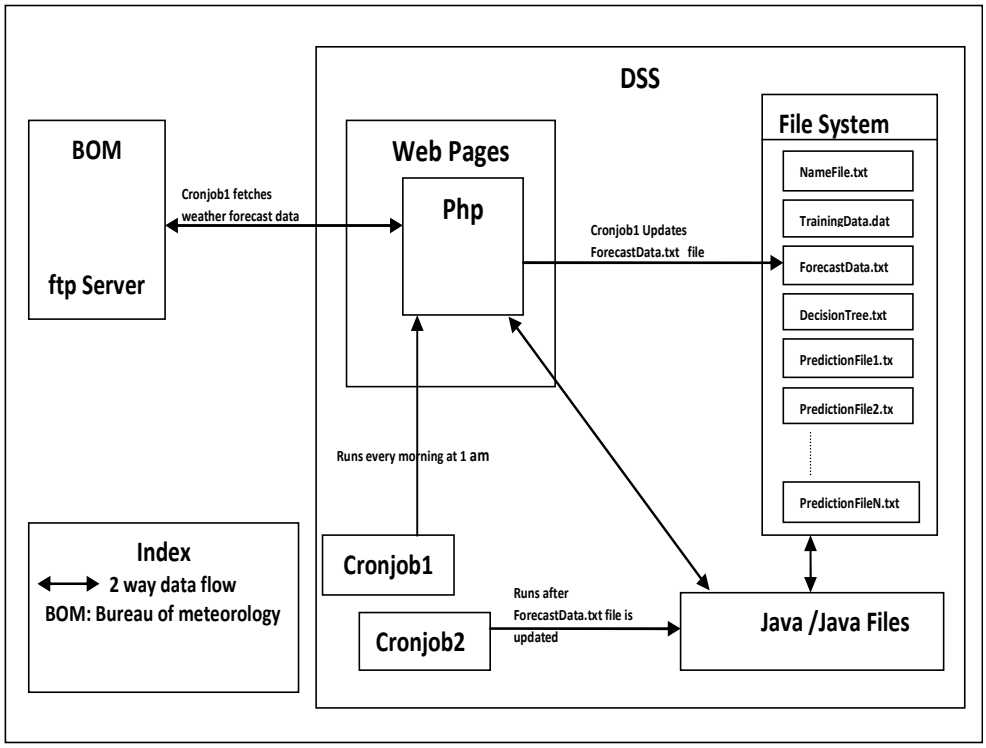

Figure 6: Conceptual Diagram Java-PHP Interaction

When the java code is invoked by PHP it connects to the file system in the DSS. To generate a decision tree, java requires two input files from the file system i) NameFile.txt and ii) TrainingData.dat.

NameFile.txt contains all information about the non-class and class attributes such as number of attributes, names of the attributes, types of attributes (numerical or categorical), possible values an attribute can have, and number of records in the training data set. TrainingData.dat contains the actual data on past records.

Every morning at 1am PHP code within the DSS automatically connects to the Bureau of Meteorology (BOM) ftp server through a scheduled cronjob1 (a cronjob is a command or a shell script which lets the users schedule jobs to run automatically at a certain time) as shown in Figure 6 and gets the weather forecast data for the next 7 days. It then replaces/overwrites only the weather data on the ForecastData.txt file in the file system, the other attributes in the ForecastData.txt file such as soil type and crop type remains same. When the contents of ForecastData.txt file is updated, the cronjob2 then executes the java code to generate the *i*th Predictionfile.txt for the *i*th farm where . The predictionfile contains the predicted values for the attribute 'crop water usage' along with all the other attributes. The prediction file will be generated for all the farms of CIA.

To generate the prediction files, java requires two input files from the file system they are i) DecisionTree.txt and ii) Forecastdata.txt. Java code first reads the DecisionTree.txt file to learn the pattern generated by the decision tree. In the second step it reads a record of the ForecastData.txt file and tests the record according to the decision tree to figure out the leaf where the record falls in. The class value of the leaf is considered as the predicted class value of the record. Thus java predicts the class value (water usage) of each record in the ForecastData.txt file.

To predict the water demand for a node (a set of a number of neighbouring farms), the water demand forecasted for the farms belonging to the node is accumulated. When a user wants to know the water demand forecast for the next 7 days for his farm, the prediction file related to his farm is displayed on the web

page. Similarly an irrigation manager can view the predictions for any individual farm, node and the whole CIA. Figure 7 shows a prediction file of a farm. Each row of the file contains all non-class attribute values and water demand prediction for the next 7 days.

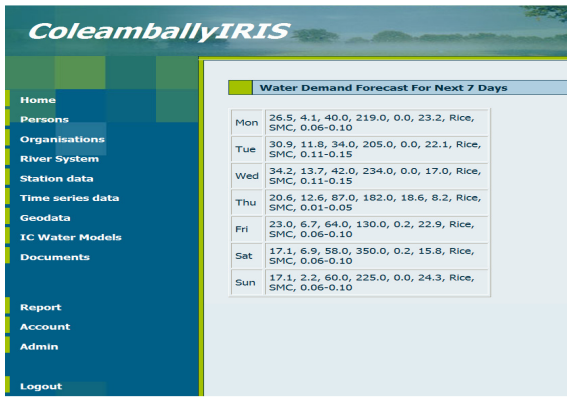

Figure 7: Irrigation water demand forecast for 7days

The water demand forecast technique in Coleambally IRIS DSS will be useful for the farmers and irrigation managers of CIA. Farmers need to order the water from CICL. Once the order is made, it often takes 7 days to get the water at the farm from the upstream. Therefore, traditionally farmers need to guess future water demand (based on their past experiences) in order to request water for the next seven days. By using our web based demand forecast technique a farmer can learn more accurate water demand for his/her farm for the next 7 days in advance. Hence, a farmer can order more accurate amount of water they require resulting in a better water savings.

5 Results and Discussion

In order to evaluate our data pre-processing technique we build two training data sets D_1 and D_2. The data set D_1 is prepared taking the Equal Distribution approach and D_2 is prepared based on our Reference Evapotranspiration Based Estimate as explained in Section 3.2. In D_1we divide the water supply made on a particular day by the number of days between this delivery and the next water delivery. Moreover, in some farms there are more than one soil types. While preparing D_1we identify the soil type of the majority area of a farm and consider this as the value of the attribute "soil type" for the records representing the farm. However, in D_2 we calculate possible water usage of each day separately based on Reference Evapotranspiration coefficient as explained in Section 3.2. Therefore, we have more accurate water usage for each record in D_2. Moreover, in order to have more accurate data, in D_2 we use only records of those farms that have a single soil type. We realise that soil type has high influence on water usage and therefore accurate information on this attribute is crucial.

We divide D_1 into two data sets; training and testing data set. We build a decision tree on the training data set and then apply the tree on the testing data set in order to check the prediction accuracy of the tree. The accuracy check is carried out using 10 folds cross validation method. This is a method of testing the accuracy of the tree by dividing the data set into 10 equal parts that are also called as folds. Nine parts of the data set are used to train the tree and the one part is used to test the tree (Han and Kamber, 2001). This process is continued 10 times so that each part of the data set is used once for testing.

D_1 data set has 6070 records in total where 607 records are used for testing in each of the ten cross validations. Similarly, D_2 data set has 1500 records where 150 records are used for testing in each of the ten cross validations. In D_1 data set there are 4570 records representing farms having multiple soil types. These records are removed from D_2 for the purity of the data set.

First we perform 10 folds cross validation on D_1 data set. We have a low accuracy of around 43%. We then added another attribute called Reference Evapotranspiration, having high correlation with water usage, in D_1 in order to check whether it increases the accuracy. However, we find that the inclusion of the additional attribute does not increase the accuracy. We also perform a 10 folds cross validation on our D_2 data set and achieve significantly high accuracy of around 74% as shown in Table 3. The result clearly indicates the effectiveness of our data pre-processing based on irrigation engineering and data mining knowledge. Moreover, the result also indicates the appropriateness of the attributes selected using three different sources namely water delivery statement, meteorological data and remote sensing processed images obtained from satellite data.

We also compare the accuracy of the decision tree based prediction model with the accuracy of a traditional approach based on actual crop evapotranspiration (ET_c). Crop evapotranspiration ET_c is calculated using crop coefficient K_c (for a crop type and cropping stage) and reference evapotranspiration (ET_o). The empirical formula to calculate ET_c is $ET_c = K_c \times ET_o$ (FAO, 1998), and this is widely used globally to estimate water demand (Hunsaker et al 2005). The crop coefficient method was developed for the farmers and irrigation managers to calculate ET_c which helps them in making irrigation management decisions (Hunsaker et al 2005). The ET_c is the same as crop water usage (Al-Kaisi and Broner, 2009).

We use the data from the year 2007/08 and 2009/10 as training data set and the data from the summer season 2008/09 as the testing data set. Currently we do not have accurate data on water usage for the winter season of 2008/09. We build a decision tree on the training data set and use the tree to predict the class values on the testing data set. We also apply the ET_c based traditional approach to estimate the water usage of the records for the testing data set. The predicted class values (obtained by the decision tree approach and the traditional approach) are then checked against the actual class values of the testing data set.

Decision tree approach achieves significantly better accuracy than the accuracy of the ET_c based tradition approach. Table 4 shows a comparison among the actual water usage, water usage predicted by the decision tree approach and water usage predicted by the ET_c base approach for all the 22 nodes of CIA for the summer cropping season of 2008-09.

Folds	Total 6070 records using a pre-processing approach.			Total 1500 records using our pre-processing approach		
	Correctly classified records	Incorrectly classified records	Accuracy Percentage	Correctly classified records	Incorrectly classified records	Accuracy Percentage
1	186	421	31	111	39	74
2	212	395	35	108	42	72
3	176	431	29	107	43	72
4	254	353	42	110	40	73
5	302	305	50	118	32	79
6	236	371	39	110	40	73
7	311	296	51	110	40	73
8	289	318	48	109	41	73
9	337	270	56	112	38	75
10	303	304	50	108	42	72
Average:			**43**			**74**

Table 3: 10 folds cross validation on data sets based on two data pre-processing methods

Both demand forecast models are applied on all the farms of CIA to obtain the water demand for a whole cropping season. Finally the water requirement for each node is calculated by adding the water demand predicted for the farms belonging to the node.

Table 4 indicates that the predicted water usage values by decision tree approach are very similar to the actual water usage values. The prediction of the ET_c based approach is not as similar to the actual water usage as the prediction of decision tree based approach for the 22 nodes of the CIA. However, in few nodes such as "Coly 7", "Bundure_Main" and "Bundure 7_8", the actual water usage is significantly lower than the water usage predicated by the decision tree method. This is because only a few farms of the nodes were irrigating during the season. Moreover, the farms stopped irrigation for some reason half way through the season as it is evident from the water delivery statement. Moreover, "Coly 10" does not have any irrigation for the cropping season. We calculate the accuracy for each node as follows.

$$Accuracy = 1 - \frac{|Actual - Predicted\ Water\ Usage|}{Actual\ Water\ Usage} \times 100\%.$$

We find the average accuracy of the decision tree based approach and the traditional ET_c based approach as 89% and 74.5%, respectively. For the accuracy calculation we exclude the four exceptional nodes (Coly 7, Bundure_Main, Bundure 7_8 and Coly 10) where irrigation was not carried out for the complete cropping season. More interestingly out of 18 nodes (that were irrigated for the complete season) only one node ("Coly 11") has the water usage prediction made by our decision tree approach worse than the prediction made by the traditional ET_c approach.

Node	Actual Water Usage (ML)	Predicted Water Usage	
		Decision Tree (ML)	ET_c(ML)
Coly 1_2	407	344	284
Coly 3	1292	1103	777
Coly 4	800	746	570
Coly 5	879	945	666
Coly 6	4359	4158	3235
Coly 7	82	220.5	157
Coly 8	785	802	875
Coly 9	4501	4297	3232
Coly 10	0	0	0
Coly 11	2262	2877.5	2264
Tubbo	696	630	444
Boona 1	1201	1069	791
Boona 2	418	372	259
Boona 3	2438	2101	1652
Yamma Main	4299	3732	3098
Yamma 1	3333	3364	3085
Yamma 2_3_4	2926	3045	2772
Bundure Main	87	493	419
Bundure 3	763	768	653
Bundure 4	1597	1058	897
Bundure 5_6	961	798	677
Bundure 7_8	133	378	268.5

Table 4: Comparison of actual and predicted water usage made by decision tree and traditional method for all 22 nodes.

6 Conclusion

The main contributions of the paper are preparation of a data set, pre-processing of the data set, implementing a decision tree based demand forecasting technique, development of a web based decision support system for managers, farmers and researchers, and carrying out of necessary experiments. We discuss the contributions one by one as follows.

We prepare a data set by carefully selecting attributes from three different sources namely water delivery statement, meteorological data obtained from our weather stations installed in the study area, and remote sensing pre-processed images obtained from satellite data. For example, we obtain meteorological data on some attributes such as solar radiation, temperature and wind speed from our weather stations. We also obtain soil type and crop type data from pre-processed satellite images. Moreover, we obtain actual water usage data from Water Delivery Statement available from CICL. We carefully take irrigation engineering and data mining requirements into consideration for selecting relevant attributes that have high influence on water usage/demand.

Since we prepare the data set from different sources we do not have all data in a desired format. For example, while we have average meteorological data for each day we do not have any information on daily water usage on a farm. In fact farmers typically do not irrigate a farm on a daily basis. From CICL water delivery statements we only learn how much water was delivered to a farm during each delivery. Once delivered a farmer can store the water in on-farm supply channels and irrigate according to his/her requirement. There is no information on the actual daily water usage by a farm. However, in order to build a useful decision tree we need daily data on weather parameters and water usage so that the tree can extract meaningful the relationship between them.

We consider a few options in order to estimate the actual daily water usage of a farm. We finally devise a technique to determine daily water usage based on the reference evapotranspiration. We use the proportion of the reference evapotranspiration of a day to the total reference evapotranspiration of the days between two water deliveries in order to estimate the possible water usage by the day. To the best of our knowledge this is a novel approach to estimate water usage in order to pre-process a data set for data mining purpose.

We then apply a decision tree based water demand forecasting approach. We also compare the approach with traditional water demand forecasting technique that is globally used by the irrigations managers and engineers.

We incorporate the decision tree based demand forecasting technique into our web based Decision Support System (DSS) so that managers and farmers can get more accurate future water requirement information from the web. Our DSS automatically collects information from the weather stations and Bureau of Meteorology (BOM). It also automatically prepares a decision tree from the processed training data set and applies the knowledge on the future data set in order to predict future water requirement. Our DSS uses PHP, java and cronjob in a combination to perform the task.

We carry out necessary experiments in order to evaluate the effectiveness of our data pre-processing approach. Our experimental results indicate a significant improvement of accuracy in water demand forecasting based on the proposed pre-processing technique when compared to the accuracy obtained from other possible techniques. Based on a 10 fold cross validation we obtain 74% accuracy on our pre-processed data set compared to 43% accuracy on the other pre-processed data set.

Moreover, we compare the decision tree based future water requirement prediction approach with a traditional evapotranspiration based technique. We experiment on all 22 nodes (made of all 473 farms) of our study area and discover that while the traditional approach has 74.5% accuracy our decision tree based technique has 89% accuracy. Our approach obtains better prediction in all except one node.

Farmers need to order water depending on their requirements. Often it takes around 7 days to get water at the farm gate after the order was made, due to propagation delay from the upstream to the farm gate. Therefore, a farmer needs to estimate the requirement of the next 7 days and request the water in advance in order to get it on time. Currently there is no reliable scientific tool available to the farmers to make an accurate estimate of future water requirements. Therefore, a farmer estimates the future water requirement purely based on his/her experience. In most of the cases they either heavily overestimate or underestimate the water requirement having adverse effect on crop production. Hence, the data mining based tool developed in the paper is crucial for the farmers to make the maximum use of limited water resource.

7 References

Al-Kaisi, M.M. and Broner, I. (2009): Crop Water use and Growth Stages, Colarado State University, leatlet no.4.715

Alvisi, S., Franchini, M. And Marinelli, A. (2007): A short-term, pattern-based model for water-demand forecasting, *Journal of Hydroinformatics*, **9**(1), 35-50.

Bontemps, C. And Couture, S. (2002): Irrigation wtar demand for the decision maker, *Environment and Development Economics,* **7:**643-657.

Coleambally Irrigation Company Limited (2010): Annual Compliance Report.

Douglas J. Hunsker., Paul J. Pinter Jr. and Bruce A. Kimball. (2005): Wheat basal crop coefficients determined by normalized difference vegetation index, Irrig Sci, **24,** 1-14.

Han, J., & Kamber, M. (2001): Data Mining: Concepts and Techniques, A Horcourt Science and Technology company, 525 B Street, Suite 1900, San Diego, CA 92101-4495, USA.

Hoogenboom, G., Paz, O.J., Salazar, Melba. And Garcia A.G. (2009): Agriculture Irrigation Water Demand Forecast, Procedures for Estimating Monthly Irrigation Demands, http://www.nespal.org/sirp/waterinfo/state/awd/AgWaterDemand_IrrAmt_Detail.htm, accessed on 19/05/2011

Hu, X. (2003): DB-HReduction: A data Preprocessing Algorithm for Data Mining Applications, *Applied Mathematics Letters*.

Islam, M. Z. (2010): EXPLORE: A Novel Decision Tree Classification Algorithm, *the 27th International Information Systems Conference, British National Conference on Databases,* June 29- July 01, 2010, Dundee, Scotland.

Islam, M. Z., Barnaghi, P. M. and Brankovic, L.(2003): Measuring Data Quality: Predictive Accuracy vs. Similarity of Decision Trees, *In Proceedings of the 6 th International Conference on Computer & Information Technology (ICCIT 2003),* Dhaka, Bangladesh, **2:** 457-462.

J.Ross Quinlann. (1993) C4.5: *Programs for machine Learning*.Morgan Kaufmann Publishers, San Mateo, California, USA.

J.Ross Quinlann. (1996): Learning Decision Tree Classifiers, *ACM Computing Surveys,* **28:**1

Jackson, T. (2009): An appraisal of the on-farm water and energy nexus in irrigated agriculture, Ph.D. thesis, Charles Sturt university, Wagga Wagga, Australia.

Khan, S., Rana, T., Dassanayake, D., Abbas, A., Blackwell, J., Akbar, S., and Gabriel, H. F. (2009): Spatially Distributed Assessment of Channel Seepage Using Geophysics and Artificial Intelligence, *Irrigation and Drainage* **58**: 307 – 320.

Miksovsky, P., Matousek, K. And Kouba, Zdenek (2002): Data Pre-Processing Support for Data Mining, *IEEE SMC*.

Pulido-Calvo, I., Roldan, J., Lopez-Luque, R. and Gutierrez-Estrada, J.C. (2003): Demand Forecasting for Irrigation Water Distribution Systems. *Journal of Irrigation and Drainage Engineering* **129**(6):422-431.

Pulido-Calvo, I. and Gutierrez-Estrada, J.C. (2009): Improved irrigation water demand forecasting using soft-computing hybrid model. *Biosystems Engineering,* 102, 202-218.

Shichao Zhang, Chengqi Zhang and Qiang Yang (2003): Data preparation for data mining, *Applied Artificial Intelligence*, **17**:5-6, 375-381.

Smith, M. (2000): The application of climatic data for planning and management of sustainable rainfed and irrigation crop production. *Agricultural and Forest Meteorology,* 102, 99-108.

Zhou, S.L., McMohan, T.A., Walton, A. and Lewis, J. (2002): Forecasting operational demand for an urban water supply zone, *Journal of Hydrology*, **259**, 189-202.