Evaluating the Performance of Several Data Mining Methods for Predicting Irrigation Water Requirement

Mahmood A. Khan [1], Md Zahidul Islam [2,3], Mohsin Hafeez [1,4]

[1] School of Environmental Sciences, Charles Sturt University, Wagga Wagga 2678, NSW, Australia
[2] School of Computing and Mathematics, Charles Sturt University, Bathurst 2795, NSW, Australia
[3] Centre for Research in Complex Systems (CRiCS), Charles Sturt University, Bathurst 2795, NSW, Australia
[4] GHD Pty Ltd, Brisbane 4000, QLD, Australia

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

Abstract

Recent drought and population growth are planting unprecedented demand for the use of available limited water resources. Irrigated agriculture is one of the major consumers of fresh water. Huge amount of water in irrigated agriculture is wasted due to poor water management practices. To improve water management in irrigated areas, models for estimation of future water requirements are needed. Developing a model for Irrigation water demand forecasting based on historical data is critical to effectively improve the water management practices and maximise water productivity. Data mining can be used effectively to build such models. Data mining is capable of extracting and interpreting the hidden patterns from a large amount of hydrological data. In recent years, use of data mining has become more common in hydrological modelling.

In this paper, we compare the effectiveness of six different data mining methods namely decision tree (DT), artificial neural networks (ANNs), systematically developed forest (SysFor) for multiple trees, support vector machine (SVM), logistic regression and the traditional Evapotranspiration (ET_c) methods and evaluate the performance of these models to predict irrigation water demand using pre-processed dataset. The pre-processed dataset we use in this study and SysFor were never used before to compare with any other classification techniques. Our experimental result indicates SysFor produces the best prediction with 97.5% accuracy followed by decision tree with 96% and ANN with 95% respectively by closely matching the predictions for water demand with actual water usage. Therefore, we recommend using SysFor and DT models for irrigation water demand forecasting.

Keywords: Irrigation water demand forecasting, Data mining, Decision tree, ANN, Multiple trees and Water management.

1 Introduction

Water scarcity is rapidly becoming a major issue for many developed and developing countries of the world,

which is a serious threat and leads to emergence of food crisis (IWMI 2009). As the scarcity of the water increases, the demand for managing available water resources becomes crucial. In particular, a recent drought in Australia has made prominent the need to manage agriculture water more wisely. It is reported that, more than 70% of available water in Australia and 70% to 80% of water Worldwide is currently being used by irrigated agriculture (Khan et al. 2009, Khan et al. 2011, IWMI 2009). Due to recent drought, climate change, population growth and increasing demand for domestic and industrial water requirement, preserving sufficient amount of freshwater for agricultural production will become increasingly difficult. Since all the existing water resources are fully utilised and drawing of more water is impracticable, therefore the best alternative is to increase the water productivity (Khan et al. 2011). Studies report that, the water delivered for irrigation is not always efficiently used for crop production, on an average 25% of water is wasted due to inefficient water management practices (FAO 1994, Smith 2000).

In order to improve water management and maximise water productivity application of various hydrological and data driven models using data mining methods have become very essential. In the current situation, models to predict future water requirements based on data mining techniques can be useful. Ullah et al. (2011) suggests that, to developing a model for water demand forecast, it is essential to understand the behaviour of the irrigation system in the past, the current land use trends and the behaviour of future hydrological attributes such as (rainfall, Evapotranspiration, seepage, etc.). Having an accurate and reliable Irrigation water demand forecasting model based on hydrological, meteorological and remote sensing data can provide important information to agriculture water users and managers (Pulido-Calvo et al. 2009, Zhou et al. 2002, Alvisi et al. 2007).

Recently, data mining techniques are increasingly being applied in the field of hydrology for developing models to predict various hydrological attributes such as rainfall, pan evapotranspiration, flood forecasting, weather forecasting etc (Pulido-Calvo *et al.* 2003). However, these techniques are not used for irrigation water demand forecasting. Knowledge discovery from any data set can be obtained through data mining. It discovers new and practically meaningful information from large datasets. Unlike any typical statistical methods, data mining techniques explores interesting and useful information without having any pre set hypotheses. These techniques are more powerful, flexible and capable

of performing investigative analysis (Olaiya et al. 2012). Zurada et al. (2005) says, data mining uses a number of analytical tools such as decision trees, neural networks, fuzzy logic, rough sets, and genetic algorithms to perform classification, prediction, clustering, summarisation, and optimisation. The most common tasks among these are classification and prediction which we carryout in this study.

The aim of this study is to explore and compare the effectiveness of accuracies of different data mining models on predicted water usage. We build models based on five data mining techniques namely decision trees, artificial neural networks, systematically developed forest (SysFor), support vector machine, logistic regression, and traditional ET_c based method. To best of our knowledge SysFor is compared with other classification techniques for the first time.

To develop an effective irrigation water demand forecasting model using data mining techniques adequate historical data for the attributes having high influence on water usage are required. We use the dataset which was collected from three different sources and pre-processed by Khan et al. (2011). The data pre-processing was carried out using a novel approach called Reference Evapotranspiration Based Estimate, which is based on Reference Evapotranspiration (ET_c), a comprehensive explanation can be found in Khan et al. (2011).

Once the models are built, we use the models to predict the water requirements for the unseen data. Our experimental results indicate a minor difference in the prediction accuracies of different data mining techniques. However, among the five different techniques/models the prediction performance of multiple decision tree technique Sysfor is found to be the best followed by Decision Tree and ANN.

This paper is organised as follows, section 2 describes the methods/models used in this study, followed by the description of study area and dataset in section 3. Experimental results are explained in the Section 4, Section 5 concludes the paper with some suggestions for future work.

2   Description of methods

All the methods/techniques used to predict water demand forecast in this study are well known and well established. Therefore, we explain only the basic functionalities of each method, without explaining the mathematical descriptions of the underlying algorithms. For more information relating to any specific algorithm on decision tree, artificial neural networks, support vector machine, systematically developed forest (SysFor) and logistic regression refer to (Quinlan 1993, Islam 2010, Khan et al. 2011; Cancelliere et al. 2002, Yang et al. 2006, Han & Kamber 2001; Vapnik 1995; Islam & Giggins 2011; Christensen, R. 1997). We explain the methods one by one as follows.

2.1   Decision Tree (DT)

Decision trees are a powerful tool for data classification. Decision tree learns from the training dataset and apply the learned knowledge on the testing dataset to find the hidden relationships between the classifying (class) and classifier (non class) attributes. A class attribute is an attribute of 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). By using the rules generated by a decision tree the relationship between the attributes of a dataset can be extracted. Each rule represents a unique path from the root node to each leaf of the tree.

Decision trees are made of nodes and leaves, as shown in Figure 1 where each node in the tree represents an attribute and each leaf represents the value for the records belonging to the leaf (Khan et al. 2011, Han & Kamber 2001). The concept of information gain is used in deciding the best suitable attribute for a node. The functionality of the decision tree is based on C4.5 algorithm (Quinlan 1993).
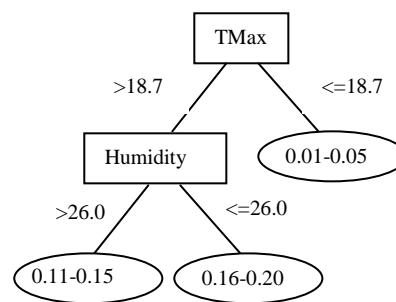
**Figure1: An Example of a Decision tree generated from our dataset**

2.2.   Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) is a data processing and classification model that is inspired by the biological neural network. ANN learns the non-linear relationships, trends and patterns from training dataset and uses the knowledge for predicting the class values of unseen datasets (Cancelliere *et al.* 2002, Yang *et al.* 2006).

Interconnection strengths known as weights are used to store the gained knowledge. Weights of the neurons in ANN are computed during the training process. Based on the nature of the datasets an appropriate network can be selected, where a user/data miner can choose number of layers and number of nodes in each layer of the network. In hydrological modelling most ANNs are trained with single hidden layer (Dawson & Wilby 2001, de Vos & Rientjes 2005) as reported by Wu *et al.* (2010). The ANN model is based on error minimisation principle. Training of the model can be carried out in two ways; supervised and unsupervised learning (Craven & Shavlik 1998, Han & Kamber 2001).

One of the most popular and commonly used ANN architectures is multilayer feed-forward neural network as shown in Figure 2, which is also called as multilayer perceptron (Muttil & Chau 2006). In a multilayer perceptron network there is an input layer, an output layer and one or more hidden layers. These layers extract patterns from a dataset and use the learned patterns to predict class values of new records. The nodes in the input layer pass the processed information to the computational nodes in a forward direction (Wang *et al.*

2009). The hidden layer is also responsible for resolving the nonlinearity between the input and output attributes of the data set (Ambrozic & Turk 2003, Cancelliere *et al.* 2002, Safer 2003).
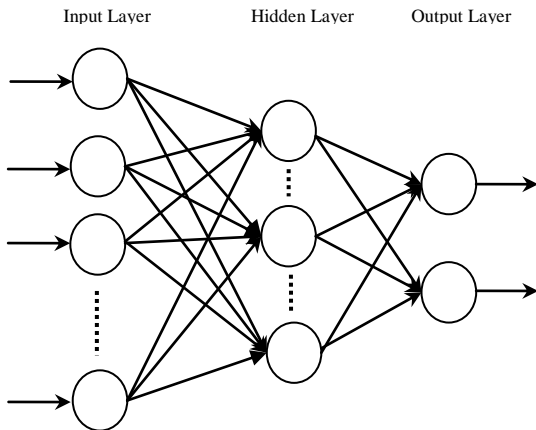
Figure 2: Architecture of three - tier feed forward neural network.

### 2.3. Systematically Developed Forest of Multiple Trees (SysFor)

SysFor is a multiple tree building technique based on the concept of gain ratio. This technique is developed by Islam & Giggins (2011). The purpose of building multiple trees is to gain better knowledge through the extraction of multiple patterns. We explain this technique in a step by step fashion.

In the first step, a set of good attributes and their split points are identified based on user defined goodness (gain ratio) and separation values. Islam & Giggins (2011) says, a numerical attribute can be chosen more than once within the set of good attributes, if it has higher gain ratios with different split points that are not close to each other. After the set of good attributes are selected and if the size of the good attributes is less than a user defined number of tree, then in the next step (step 2) SysFor builds the tree using each good attribute as the root attribute of the tree, and build as many trees as number of good attributes. Else it builds user defined number of trees from the set of good attributes as the root attribute.

If the number of trees build in this step are still less than the user defined number of trees, then SysFor in the next step (step 3) build more trees until user defined number is met by using alternative good attributes at the next level of the tree i.e. at level 1 of the tree generated in the previous step (step2). In this step (step 3) the algorithm first uses the root attribute of the first tree built in step 2 in order to split dataset into horizontal partition. The algorithm, then selects a new set of good attributes, their respective split points and a set of gain ratios for each horizontal partition. Based on these set of good attributes the algorithm builds a tree from each partition and the trees are joined by connecting their roots (at level 1) to the root (at level 0) of first tree build in step 2. This process of building more trees continues until user defined number of trees are generated/build. Example trees generated in SysFor are shown in Figure 3a, 3b.

After Systematic forest of multiple trees is generated as to predict the class values of unseen records we follow

voting system proposed by Islam and Giggins (2011) called SysFor Voting-2. In this voting system, we find all the leaves from all the trees the record falls into. Then the leaf with highest accuracy is determined (based on maximum number records with same class values to total number of records) and finally the majority class value of the leaf is chosen as the predicted class value of the record.
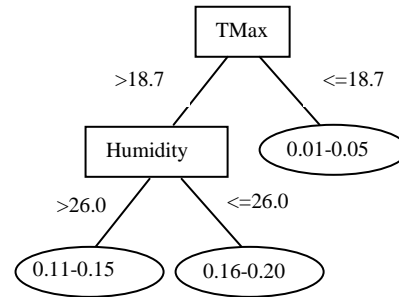
Figure 3a: Tree generated in SysFor based on first good attribute
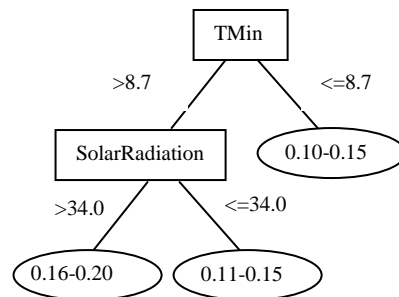
**Figure 3b: Tree generated in SysFor based on second good attribute**

2.4. Support Vector Machine (SVM)

Support vector machine is a state of the art neural network methodology based on statistical learning (Vapnik 1995, Wang et al. 2009). An SVM is an algorithm for maximizing a particular mathematical function with respect to a given dataset. The basic concepts behind the SVM algorithm are i) the separating hyperplane, ii) the maximum-margin hyperplane, iii) the soft margin and iv) the kernel function. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification. In general, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class as shown in Figure 4 which exhibits the basic concept of support vector machine. From Figure 4 is it can be seen that the optimal hyperplane separates the positive and negative points from the dataset with a maximum margin, indicating the maximum distance to hyperplane from closest positive and negative data points.

2.5. Evapotranspiration (ET_c) based Prediction

ET_c can be broadly defined as crop water usage. 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 $ET_c= K_c$ x ET_o (FAO 56), and

this is commonly used globally to estimate water demand. The crop coefficient method was developed for the agriculture users to calculate ET_c which helps them in making irrigation management decisions.
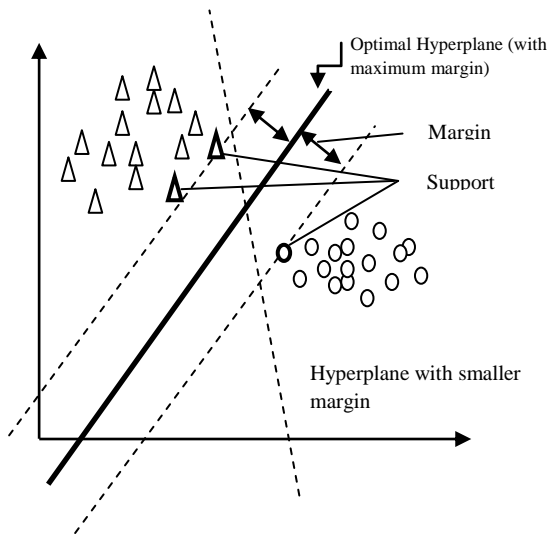
Figure 4: Basic concept of support vector machine

2.6.   Logistic Regression

The main goal of logistic regression model is to predict the label *t* of a new given data point *x* based on the learning from the training data set. Logistic regression can be of two types 1) Simple Logistic Regression and ii) Multiple Logistic regression. Simple logistic regression is used to predict the class value, given it is categorical and has only two possible outcomes such as (male/female). Whereas, the multiple logistic regression can be used to predict the class value consisting of three or more possible outcomes.

Logistic regression is a capable probabilistic binary classifier (Christensen 1997). A logistic regression model helps us assess probability from which the outcomes will be chosen.

It is evident from the literature that the logistic regression is used extensively in numerous disciplines such as, in the field of medical and social sciences, marketing applications etc (Pearce& Ferrier, 2000). Zurada 2005, states that logistic regression models are designed to predict one class value at a time and they are assumed as simplest feed forward neural networks containing only two layers input and output.

3   Study area and Dataset

In this study, Coleambally Irrigation Area (CIA) is selected as our study area. CIA is one of the most modernized irrigation areas in the Murray and Murrumbidgee river basins of Australia. CIA is situated approximately 650km south-west of Sydney in the Riverina District of New South Wales which falls under lower part of Murrumbidgee River catchment as shown in Figure 5. CIA contains approximately 79,000ha of intensive irrigation area and 325,000ha of the Outfall District area, supplying water to 495 irrigation farms (CICL, 2011). Because of the recent drought in the last decade, there is a significant decline in the average water allocation to the farmers of CIA. Due to declining water

allocation and changing weather patterns, CIA requires new management measures for water use efficiency and increase water productivity.

The data for the years 2007/08 and 2009/10 is used to train the models and the data for summer season of 2008/09 is used to test the models. We use the same dataset which was collected from three different sources namely Water delivery statements, Meteorological data, and Remote sensing data and pre-processed by Khan et al. (2011) consisting of 1500 records. Khan et al. (2011), claims the dataset was pre-processed using a novel method which is based on Reference Evapotranspiration (ET_o) and is the combination of knowledge in irrigation engineering and data mining. The main goal to pre-process that dataset was to estimate daily crop water usage more accurately based on the data collected from water delivery statements.

The dataset consist of historical data on weather parameters such as Maximum and Minimum temperature, wind speed, humidity, rainfall and solar radiation in combination with soil type, crop type and crop water usage. Attributes crop type and soil type are categorical and the rest are numerical. In our experiments, we consider crop water usage as the class attribute and the rest as non-class attributes, also crop water usage is considered as a categorical attribute.

## 4   Experimental Results

The main purpose of this experiment is to compare the prediction performances of different data mining models on water demand forecasting.

We first built a decision tree from our training dataset to extract the relationship between the non-class and class attributes. We implement C4.5 algorithm to generate a decision tree. C4.5 takes a divide and conquers approach to build a decision tree from a training dataset using the principle of information gain (Quinlan 1993). Here we divide our dataset into two parts training and testing, the tree is built on training dataset and applied on testing dataset to check the prediction accuracy of unseen records.

In this study, an ANN is built using the three tier feed-forward architecture with back propagation. In order to build an ANN, we divide the datasets into three parts; 70%, 20% and 10% for training, validating and testing, respectively. Training of the network is performed using two different network topologies, firstly by using 1 hidden layer having 8 nodes, and secondly by using 1 hidden layer having 6 nodes. Both the networks are trained for 30000, 50000 and 70000 learning iterations. The network produced by 1 hidden layer with 8 nodes for 30000 learning iterations produces better results. The ANN is built using EasyNN plus V14.0 software (available from http://www.easynn.com/).

We also build SysFor on our dataset, by considering user defined number of trees to be 5 and follow SysFor voting 2 for predicting the unseen records.

Finally we train and test SVM and Logistic regression using WEKA 3.6.2 which is available at http://www.cs.waikato.ac.nz/~ml/weka/ and very popularly used tool for performing different data mining tasks.
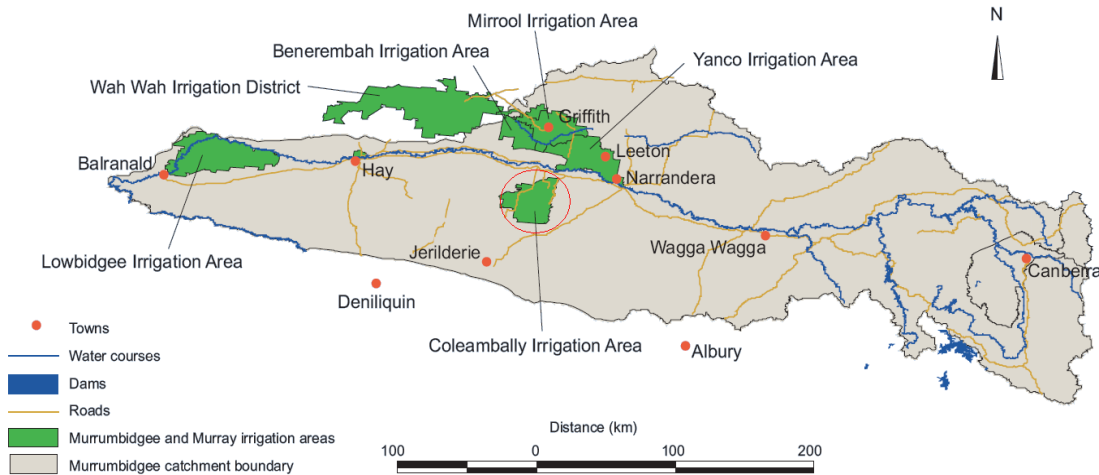
Figure 5: Location of Coleambally Irrigation Area and Other Major Irrigation Areas in Murrumbidgee Catchment

The performance evaluation of the models is carried out by comparing the prediction accuracies. The prediction accuracy check is performed using a 3 fold cross validation method. This is a method of testing the accuracy by dividing the dataset in three equal parts also called as folds, where two parts of the dataset are used for training and the third part is used for testing. This process is continued 3 times so that each part of the dataset is used once for testing. Table 1 displays the prediction accuracies of all the models used in our experiment.

Table 1 indicates that the performance of multiple decision tree technique Sysfor is better among all the other techniques, followed by decision tree and SVM. SysFor records 78% prediction accuracy while DT and SVM exhibit an accuracy of 74% and 64% respectively. The accuracy of ANN and logistic regression were recorded low. We also compare the accuracies of the experimented models with the accuracy of traditional approach which is based on actual crop evapotranspiration (ET_c).

Folds	Model				
	DT	ANN	SysFor	SVM	Logistic Regression
1	72.6	59.3	75.5	63.2	57.1
2	74.5	60.7	83	62.1	53.7
3	73.8	62	77.9	67.1	56.7
Average (%)	74	61	78	64	56

Table 1: Prediction accuracies of different models based on 3 folds cross validation

Apart from accuracy test we also compare the closeness of actual water consumed by the crop to the water predicted by the above mentioned models for summer season of the year 2008/09. Table 2 shows a comparison between the actual water usage, water usage predicted by the decision tree, ANN, SysFor, SVM, Logistic regression and traditional ETc based approach for all the 22 nodes of CIA.

All the models are applied on every farm of CIA to obtain the water demand for a whole cropping season. The water demand for each node is calculated by adding the water demand predicted for the farms belonging to the node. The accuracy of closeness for actual and predicted water is calculated as follows

$$Accuracy = 1 - \left(\frac{|Actual\text{-}Predicted\ Water\ Usage|}{Actual\ Water\ Usage}\right) \times 100\%$$

From Table 2 it is evident that the water demand predicted by SysFor is more closely matching the actual water consumed. The accuracy of closeness is found to be 97.5% which suggest a high closeness of prediction made by the model. The accuracy of SysFor is followed by decision tree and ANN whose closeness is found to be 96% and 95% which is also considered to be very high. However, in few nodes such as Yamma1and Boona 2 the prediction of SysFor was worse than decision tree and ANN. In majority of the nodes the performance of SVM, Logistic regression and ET_c was behind the performance of Sysfor, decision tree and ANN.

Moreover, 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 all the models. This is because only a few farms of the nodes were irrigating during the season. 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 exclude results of these nodes while calculating the accuracy of the models. In Table 2 the rows representing the above said nodes are shaded to highlight the exclusion of these nodes.

Figure 6 and Figure 7 displays the basic comparison between actual and predicted water usage. Figure 6 show the positive (predicted more) and negative (predicted less) predictions to actual water usage for all 22 nodes of CIA from all six models. It is evident from Figure 6 that the bars representing SysFor and DT are shorter for all nodes compared to the longer bars representing other models.

Node	Predicted Water Usage						
	Actual Water Usage (ML)	Decision Tree (ML)	ANN (ML)	SysFor (ML)	SVM (ML)	Regression (ML)	ET$_c$ (ML)
Coly 1_2	407	344	316	379	417	428	284
Coly 3	1292	1203	1210	1155	1278	1417	777
Coly 4	800	746	1262	759	841	931	570
Coly 5	879	945	1383	1001	1110	1228	666
Coly 6	4359	4158	3807	4464	4891	5266	3235
Coly 7	82	220.5	245	231	256	283	157
Coly 8	785	802	830	850	1084	1139	875
Coly 9	4501	4297	4394	4317	4801	5211	3232
Coly 10	0	0	0	0	0	0	0
Coly 11	2262	2877.5	3104	2581	2996	3139	2264
Tubbo	696	630	814	645.7	716	792	444
Boona 1	1201	1069	1692	1189	1323	1424	791
Boona 2	418	429	531	550	720	797	259
Boona 3	2438	2101	2268	2341	2585	2713	1652
Yamma Main	4299	3732	4542	4375	4921	4966	3098
Yamma 1	3333	3364	3100	3940	5558	5558	3085
Yamma 2_3_4	2926	3045	3207	3180	4479	4370	2772
Bundure Main	87	493	650	646	726	745	419
Bundure 3	763	768	636	798	897	901	653
Bundure 4	1597	1384	1421	1387	1560	1532	897
Bundure 5_6	961	798	660	836	935	941	677
Bundure 7_8	133	378	504	396	440	486	268.5
Coleambally Irrigation Area	**33917**	32692.5	35177	**34747.7**	41112	42753	26231

Table 2: **Comparison of water usage predicted by different models to actual water usage for all nodes of CIA**
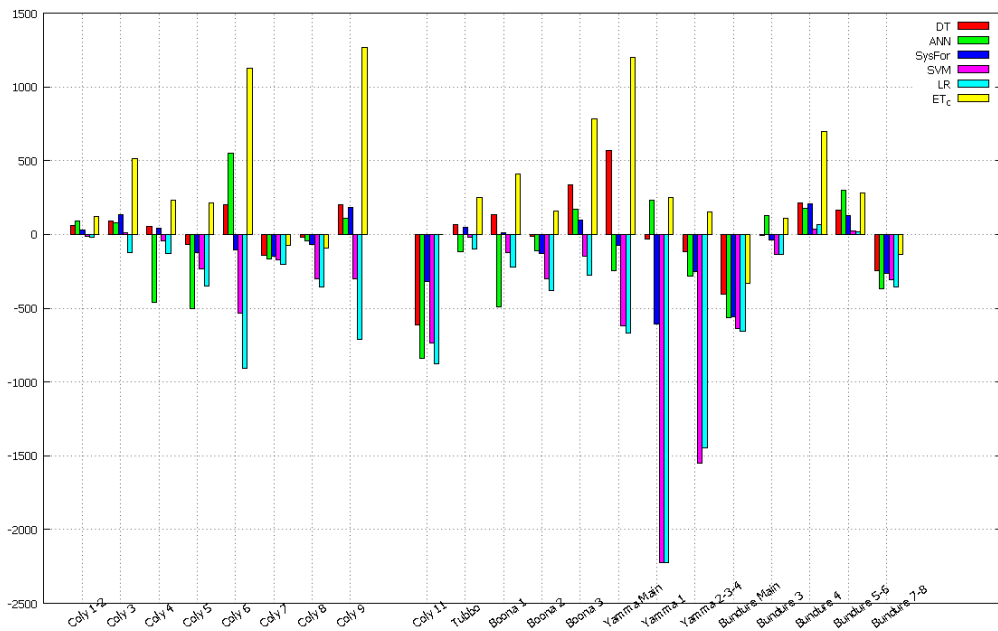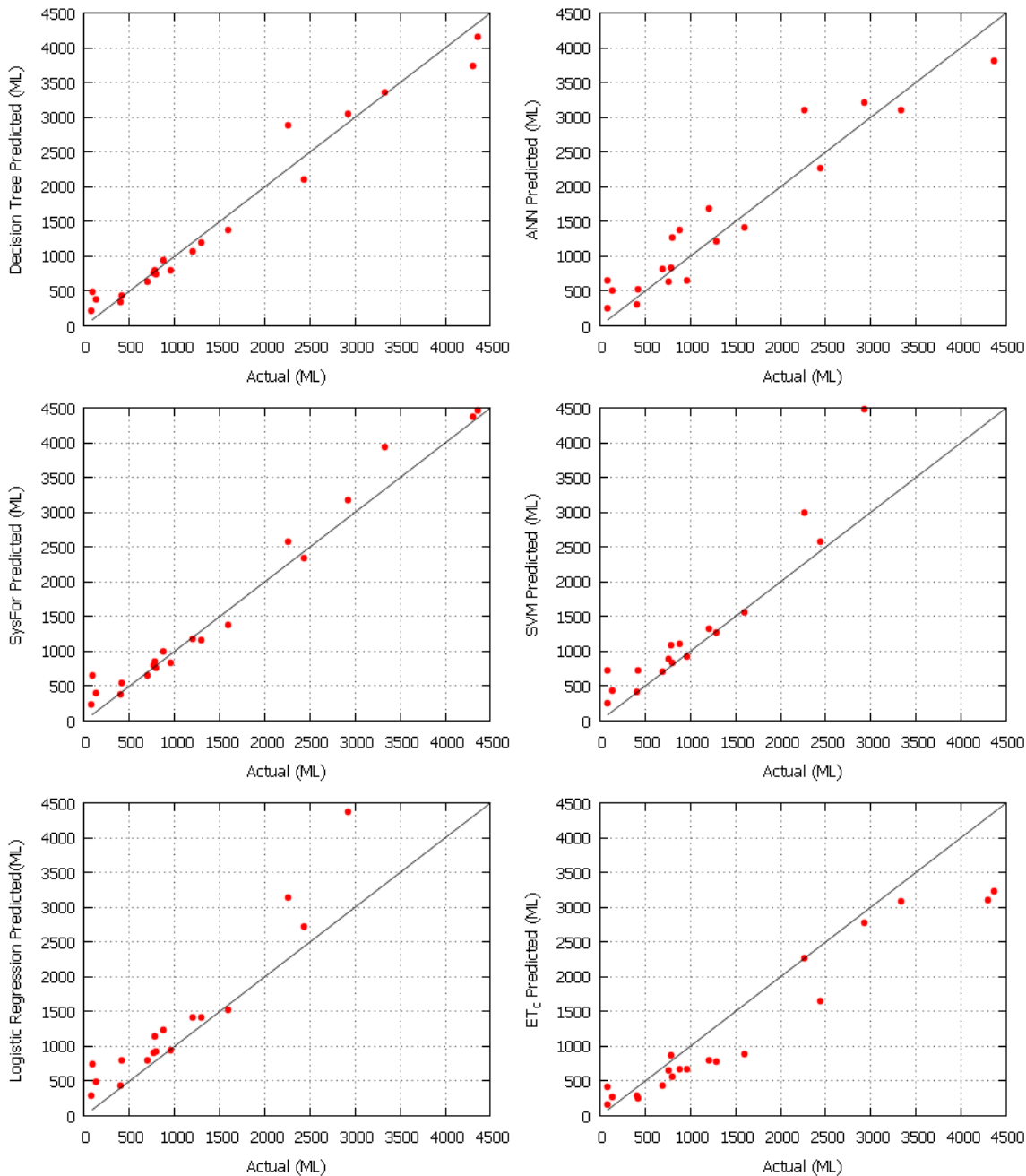

Figure 6: Positive and Negative difference between actual and predicted water usage made by different models

Figure 7: Actual Vs Predicted Water Usage made by six different models on 22 nodes of CIA

Therefore, we can say that the predictions made by SysFor and DT are close to actual water usage. Similarly, the scatter plots in Figure 7 shows the actual and predicted water usage made by all the models experimented in this study.

We also developed a web based Decision Support System (DSS) called Coleambally IRIS which consists of a database and collection of various models. Users (farmers and irrigation managers) access various data from DSS including water predictions made by our model as shown in Figure8. Based on our previous study we incorporated Decision Tree model in our DSS for predicting future water requirements. By using demand forecast results users will learn the water requirement for their particular farm for 7days in advance and can order the exact amount of water they require, this will increase the percentage of water savings and improve water use efficiency.

5   Conclusion

This study compares the effectiveness and performances of several data mining techniques such as decision tree, ANN, SysFor, SVM and logistic regression in predicting irrigation water demand. The novelty of this study is comparison of SysFor with other classification techniques which to our knowledge was carried out for the first time, and the application of pre-processed dataset on different classifier models.

Our experimental results indicate a minor difference in the prediction accuracies achieved by different data mining techniques mainly SysFor, Decision tree and ANN. Computational results demonstrate that based on 3 folds cross validation method multiple decision tree technique SysFor produce the best prediction accuracy of

78% followed by decision tree and SVM with 74% and 64% respectively.

We also compare the prediction accuracies of the models with the actual water consumed by the crop. The closeness of prediction accuracy of SysFor performs the best with 97.5% followed by decision tree with 96% accuracy. Interestingly, ANN performs better than SVM by closely predicting the water demand to actual water used with 95% accuracy. The accuracy predictions made by SVM, logistic regression and traditional ET_c method are found to be 78%, 75% and 77% respectively.

Therefore, from the above results we recommend that SysFor, decision tree and ANN techniques are most suitable for predicting irrigation water demand. By developing and implementing a demand forecasting model using these techniques the farmers and irrigation managers of CIA can learn the future water requirement in advance accurately. Hence, this tool is crucial for effectively improving existing water management practices and maximising water productivity. Although the results obtained from this study are more significant for predicting water demand, the limitation would be use of less number of influential attributes in the dataset. This can be further improved by adding more attributes having high influence on crop water usage such as seepage, soil moisture, etc. In addition it would be interesting to explore the influence of cropping stage on crop water use. Furthermore, based on our results from this study we plan to incorporate SysFor model into our DSS to make the water predictions more accurate and reliable.
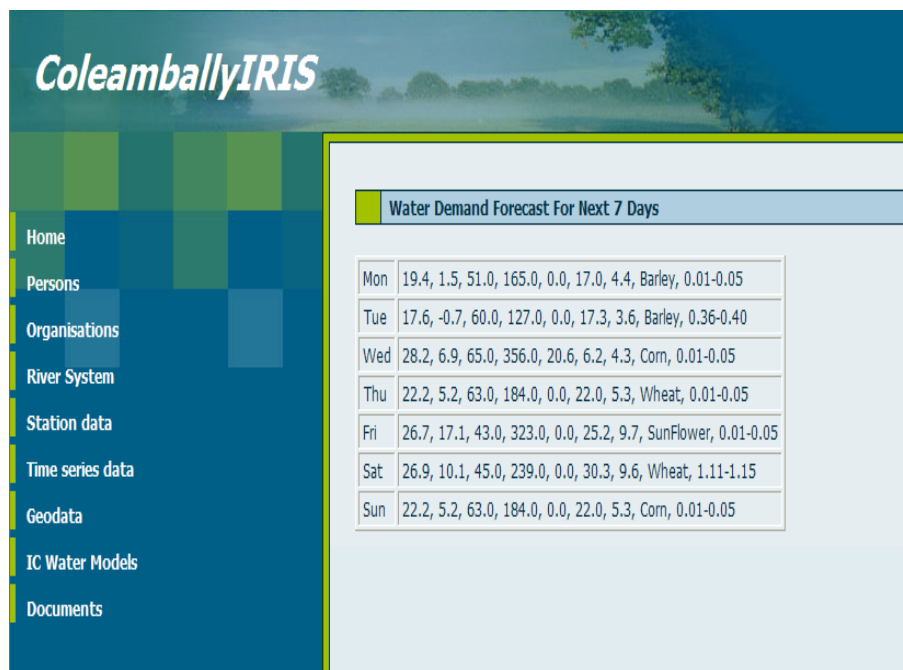

Figure 8: Irrigation water demand forecasting for 7 days

6   References

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.

Ambrozic, T & Turk, G. (2003): Prediction of subsidence due to underground mining by artificial neural networks. *Computers and Geosciences,* 29, 627–637

Cancelliere, A., Giuliano, G., Ancarani, A. & Rossi, G. (2002): A Neural Networks Approach for Deriving Irrigation Reservoir Operating Rules. *Water Resource Management,* 16, 71-88

Christensen, R. (1997): Log-Linear Models and Logistic Regression. Springer.

Coleambally Irrigation Company Limited (2011): Annual Compliance Report.

Dawson, C.W. & Wilby, R.L. (2001): Hydrological modelling using artificial neural networks, *Progress in Physical Geography,* 25 (1), 80-108.

De Vos, N.J. & Rientjes, T.H.M. (2005): Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation. *Hydrology and Earth Systems Sciences,* 9, 111-126.

FAO56, FAO Irrigation and Drainage Paper, http://www.kimberly.uidaho.edu/ref-et/fao56.pdf accessed on 25/7/2012.

FAO, 1994 Water for Life. World Food Day 1994, Rome.

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

IWMI (2009): Water for a food secure world. International Water management Institute (IWMI) Strategic Plan 2009-2013.

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. and Giggins, H. (2011): Knowledge Discovery through SysFor: A Systematically Developed Forest of Multiple Decision Trees. *In*

Proceedings of the 9th Australasian Data Mining Conference (AusDM 11), Ballarat, Australia. Dec 01 - Dec 02, 2011. CRPIT, 121, 195-204.

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.

Khan, M., Islam, M. Z., Hafeez, M. (2011): Irrigation Water Demand Forecasting – A Data Pre-Processing and Data Mining Approach based on Spatio-Temporal Data, *In Proceedings of 9th Australasian Data Mining Conference (AusDm11)*. Ballarat, Australia. Dec 01-Dec 02, CRPIT, 121, 183-194.

Muttil, N. & Chau, K.W. (2006): Neural Network and Genetic Programming for Modelling Coastal Algal Blooms. *International Journal of Environment and Pollution,* Vol. **28**, N0.3-4, 223-238.

Pearce, J. and Ferrier, S. (2000): Evaluating the predictive performance of habitat models developed using logistic regression, *Ecological Modelin,* **133**(3), 225-245

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.

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

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

Safer, A. M. (2003): A Comparison of two data mining techniques to predict abnormal stock market returns, *Intelligent Data Analysis,* **7**, 3-13.

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.

Olaiya,F., Adeyemo, A. B. (2012): Application of Data Mining Techniques in Weather Prediction and Climate Change Studies, *I.J. Information Engineering and Electronic Business*, **1**, 51-59

Ullah, K, and Hafeez, M. (2011): Irrigation Demand forecasting using remote sensing and meteorological data in semi-arid regions. *In Proceedings of Symposium J-H01 held during IUGG2011*, Melbourne, Australia, July 2011, 157-162.

Vapnik, V. (1995): The Nature of Statistical learning Theory, Springer, New York.

Wang, W.C., Chau, K.W., Cheng,C.T. & Qiu, L. (2009): A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series, *Journal of Hydrology,* **374** (2009), 294-306.

Wu, C.L., Chau, K.W. & Fan, C. (2010): Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques, *Journal of Hydrology,* **389** (2010), 146-167.

Yang, L., Dawson, C.W., Brown, M.R. & Gell, M. (2006): Neural network and GA approaches for dwelling fire occurrence prediction. *Knowledge-Based Systems,* **19**, 213-219.

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

Zurada, J. and Lonial, S. (2005): Comparison of the Performnce of Several Data Mining Methods for bad Debt Recovery In The Health Industry, *The Journal of Applied Business research*, **21**, 37-54.