AN INTELLIGENT APPLICATION OF FUZZY ID3 TO FORECAST SEASONAL RUNOFF

Chandar Sahu
Lakshmi Narain College of Technology, Indore
chandar.sahu@gmail.com

ABSTRACT

Agriculture in India has a significant history. Today, India ranks second worldwide in farm output. Agriculture and allied sectors like forestry and fisheries accounted for 16.6% of the Gross Domestic Products (GDP) in 2009, about 50% of the total workforce. But most of farmers are depend on the seasonal weather conditions for agriculture issues. And if farmer's lands are captivities thus the water supply management also needs this forecast to plan water allocations for the following summer season. This paper proposes a hybrid forecasting technique that is combination of fuzzy logic and Iterative Dichotomizer (ID3) data mining algorithm. That forecast seasonal runoff on the basis of daily weather conditions

KEYWORDS
fuzzy logic; ID3; water supply; forecast

1. INTRODUCTION

Water managers are frequently required to ensure a continuous water supply to meet such demands as consumption, agriculture and the environment Information about the extent, spatial distribution and temporal variation of runoff at regional scales is essential to understand its influence on regional hydrology, as well as conservation and development of land resources. Conventional techniques of runoff measurement are useful, however in most cases such measurements are very expensive, time consuming and difficult. Therefore, rainfall-runoff models are commonly used for computing runoff.

The Global Circulation Model (GCM) output of Hadley centre (HADCM3) projected climate change data was used. Scenario for 2080 (A2 scenario indicating more industrial growth) was selected. The runoff was modeled using the Curve Number (CN) method in spatial domain using satellite derived current Landuse/cover map. The derived runoff was compared with the runoff using normal climatic data (1951-1980). The results showed that there is a decline in the future climatic runoff in most of the river basins of India compared to normal climatic runoff. However, significant reduction was observed for the river basins in the eastern region viz: lower part of Ganga, Bahamani-Baitrani, Subarnrekha and upper parts of the Mahanadi. The mean runoff reduction during 4 months (June- September) were 66 mm, 110 mm, 120 mm and 113 mm for Brahmaputra-Barak Subarnrekha, Subarnarekha and Brahmini-Baitrani basin, respectively in comparison to normal climatic runoff. Overall seasonal (June to September) runoff reduction was high for Subarnrekha basin (54.1 %). Rainfall to runoff conversion was high for Brahmaputra-Barak basin (72 %), while coefficient of variation for runoff was more for Mahanadi basin (1.88). Study indicates that eastern India agriculture will be affected due to shortage of surface water availability.

DOI : 10.5121/ijci.2013.2103

2. LITERATURE SURVEY

There are various system implemented but most of the system found for weather forecasting system. There are too few systems for forecasting of seasonal runoff. Some of seasonal runoff systems found but they are implemented using the following techniques.

- neural network
- SVM(support vector machine)
- simple fuzzy logic
- simple decision trees

The above given techniques having its own advantages and disadvantages these disadvantages are listed below.

Sr. No	Method	Description
1	Neural Network	Provide high accuracy but required more training cycle
2	SVM	Provide better accuracy but complex calculations and large amount of memory used
3	Fuzzy system	Accurate results but in opaque model
4	Decision Tree	Having not good accurate results

3. BACKGROUND

3.1 Need for Data Analysis

Analysis of data is a process of inspecting, clean-up, transforming, and modeling data with the purpose of importance useful information, suggesting conclusions, and supporting judgment making. Data analysis has multiple facets and approaches, surrounding diverse techniques under a variety of names, in different business, science, and social science domains.

Data mining is a data analysis technique that focuses on modeling and information discovery for extrapolative rather than purely expressive purposes. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses.

Predictive analytics focuses on submission of statistical or structural models for predictive forecasting or classification, while text analytics applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties of data analysis.

Data integration is a precursor to data analysis, and data analysis is closely linked to data visualization and data dissemination. The term data analysis is sometimes used as a synonym for data modeling. Data analysis is a process, within which several phases can be distinguished Data.

3.2 Cleaning

Data cleaning is an important procedure during which the data are inspected, and erroneous data are—if necessary, preferable, and possible—corrected. Data cleaning can be done during the stage of data entry. If this is done, it is important that no subjective decisions are made. The guiding principle provided by Adder (ref) is: during subsequent manipulations of the data, information should always be cumulatively retrievable. In other words, it should always be possible to undo any data set alterations. Therefore, it is important not to throw information away at any stage in the data cleaning phase. All information should be saved (i.e., when altering variables, both the original values and the new values should be kept, either in a duplicate data set or under a different variable name), and all alterations to the data set should carefully and clearly documented, for instance in a syntax or a log.[2].

3.3 Initial data analysis

The most important distinction between the initial data analysis phase and the main analysis phase, is that during initial data analysis one refrains from any analysis that are aimed at answering the original research question. The initial data analysis phase is guided by the following four questions. [2].

3.4 Quality of data

The quality of the data should be checked as early as possible. Data quality can be assessed in several ways, using different types of analyses: frequency counts, descriptive statistics (mean, standard deviation, and median), normality (skew ness, kurtosis, frequency histograms, normal probability plots), associations (correlations, scatter plots).

Other initial data quality checks are:

Checks on data cleaning: have decisions influenced the distribution of the variables? The distribution of the variables before data cleaning is compared to the distribution of the variables after data cleaning to see whether data cleaning has had unwanted effects on the data.

Analysis of missing observations: are there many missing values, and are the values missing at random? The missing observations in the data are analyzed to see whether more than 25% of the values are missing, whether they are missing at random (MAR), and whether some form of imputation is needed.

Analysis of extreme observations: outlying observations in the data are analyzed to see if they seem to disturb the distribution.

Comparison and correction of differences in coding schemes: variables are compared with coding schemes of variables external to the data set, and possibly corrected if coding schemes are not comparable.

The choice of analyses to assess the data quality during the initial data analysis phase depends on the analyses that will be conducted in the main analysis phase. [2].

3.5 Characteristics of data sample

In any report or article, the structure of the sample must be accurately described. It is especially important to exactly determine the structure of the sample (and specifically the size of the subgroups) when subgroup analyses will be performed during the main analysis phase.

The characteristics of the data sample can be assessed by looking at:

- Basic statistics of important variables
- Scatter plots
- Correlations
- Cross-tabulations[3]

Final stage of the initial data analysis

During the final stage, the findings of the initial data analysis are documented, and necessary, preferable, and possible corrective actions are taken.Also, the original plan for the main data analyses can and should be specified in more detail and/or rewritten. In order to do this, several decisions about the main data analyses can and should be made:

- In the case of non-normal: should one transform variables; make variables categorical (ordinal/dichotomous); adapt the analysis method?
- In the case of missing data: should one neglect or impute the missing data; which imputation technique should be used?
- In the case of outliers: should one use robust analysis techniques?
- In case items do not fit the scale: should one adapt the measurement instrument by omitting items, or rather ensure comparability with other (uses of the) measurement instrument(s)?
- In the case of (too) small subgroups: should one drop the hypothesis about inter-group differences, or use small sample techniques, like exact tests or bootstrapping? In case the randomization procedure seems to be defective: can and should one calculate propensity scores and include them as covariates in the main analyses [3].

4. Problem Analysis and Solution Domain

Most of the Indian people leaves in village and main earning source of these people is depends upon agriculture. But main income source of these people is depends upon seasonal rain fall and whether conditions. Thus an accurate runoff forecasting system is required. There is some other forecasting technique is available but the authenticity of these models are unknown. Thus required to compare this model performance to older system.To solve the above described problem I will propose a new method for calculation of seasonal runoff over the simple fuzzy system. In the proposed model I will combine fuzzy logic to the ID3 traditional algorithm. And generate a new system.

My complete work is performed in the below given steps.

- implementation of fuzzy runoff forecasting
- implementation of fuzzy ID3 runoff forecasting
- calculate the performance of both systems
- compare which one is best for forecasting

4.1 Overview of ID3

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric ---information gain.To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function.

4.2 Entropy

In order to define information gain exactly, we require discussing entropy first. Let's assume, without loss of simplification, that the resultant decision tree classifies instances into two categories, we'll call them P (positive) and N (negative)
Given a set S, containing these positive and negative targets, the 'entropy of S related to this Boolean classification is:

Entropy(S) = - P (positive) log2P (positive) -P (negative) log2P
(Negative)
P (positive): proportion of positive examples in S
P (negative): proportion of negative examples in S

4.3 Information Gain

As we mentioned before, to minimize the decision tree depth, when we traverse the tree path, we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best choice.We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node.
The information gain, Gain(S, A) of an attribute A,

Gain(S,A)= Entropy(S) -Sum for v from 1 to n of (|Sv|/|S|) * Entropy(Sv)
We can use this notion of gain to rank attributes and to build decision trees where at each node is located the attribute with greatest gain among the attributes not yet considered in the path from the root.

The intention of this ordering is:

1. To create small decision trees so that records can be identified after only a few decision tree splitting.
2. To match a hoped for minimalism of the process of decision making

5. CONCLUSIONS

In this paper a hybrid algorithm is proposed for mining the sample data. Hybrid algorithm is designed using fuzzy logic theory and ID3 decision tree. In this hybrid algorithm sample data is processed using fuzzy logic and the output of the fuzzy is supplied to the ID3 decision tree to generate rules from the data model. The system generated rules are used to predict seasonal runoff.

REFERENCES

[1] Dubois D. 1980. "Fuzzy Sets and Systems, Theory and Applications." Academic Press: New York.

[2] C.Mahabir et al.,"Application of fuzzy logic to forecast seasonal runoff", Department of Civil & Environmental Engineering, University of Alberta,Edmonton, Alberta T6G 2G7, Canada

[3] Wei Peng et al., "An Implementation of ID3 --- Decision Tree Learning Algorithm",Project of Comp9417: Machine Learning University of New South Wales,School of Computer Science & Engineering, Sydney, NSW 2032, Australia.weipengtiger@hotmail.com

[4] Ratna Nayak et al., "An Enhanced approach for Weather Forecasting using Neural Network."

[5] Bagging One-class Decision Tree Chen Li college of information engineering, Northwest A&F Unoiversity,Yangling, Shaanxi province,P.R.china712100ichen_0810@nwsuaf.edu.cn

[6] Rahib H. Abiyev et al., "A type-2 neuro fuzzy system based on clustering and gradient techniques applied to system identification and channel equalization."

Authors

Chandar Sahu, Lakshmi Narain College Of Technology, Indore
(e-mail id:- chandar.sahu@gmail.com)