International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Prediction and Analysis of Water Resources using Machine Learning Algorithm

Sarakutty T. K., Ravikumar K., Hanumanthappa M.

Abstract: ***Water demand prediction plays an important role in urban and environmental planning, ecological development, decision-making processes and optimum utilization of water resources. A precise water demand prediction has a key job in the forecasting, design, process, and organisation of water resources frameworks. The under stress natural resources and the ever increasing population size makes it dominant to accurately and efficiently forecast water demand in the urban area which is possible by applying data mining techniques on the huge volumes of available water data. This paper focuses on building precise predictive models for water demand prediction using support vector machine which takes care of the nonlinear changeability of water demand at diverse levels for optimal operations.***

Keywords: Data Mining, Machine Learning, Support Vector Machine.

I. INTRODUCTION

Ecological organization of water resources is very important since the economic development of several countries depend on it. Water demand prediction is a crucial factor in ecological water resource forecasting and organization which helps in finding novel water resources. There is an expanded requirement for water demand prediction in light of the fact that it can give a simulated perspective on future and recognize reasonable administration choices to adjust water supply and demand. Accurate water demand prediction is important due to the increase in population growth rates and increase in development growth in urban areas [1]. Exact family unit level figures might be utilized for accurately foreseeing future interest and in distinguishing families that are anticipated to utilize huge amounts of water later on as potential focuses for productivity checking. In any case, making these exact prescient models for utilities information can be trying because of the huge amount of information present [2]. Numerous methods are presented for predicting urban demand for water. The choice of a particular method is based on the available data. In order to predict urban water demand various methodologies like econometric prediction, end-use prediction and time series prediction are available. Deterministic or probabilistic models are the models used for prediction. Deterministic models are extensively used for water prediction.

Accomplishing the ideal expectation exactness is a difficult assignment in light of the fact that the prediction model should at the same time think about an assortment of dominant features like explanatory variables which influence water request. These incorporate socioeconomic constraints like housing density, population density, occupation, income and water levy; climate information like precipitation and temperature; protection measures and social factors like customer inclinations and propensities. While uncertainties occur during prediction; probabilistic model can be used, wherein the quantifiable uncertainties are identified by allocating probability distribution functions to all explanatory variables inside the prediction model [3]. Water consumption usage data can be collected with at most accuracy using technologies, like smart meter and level sensors. The forthcoming water demand can be modeled better with prediction techniques. The short time prediction of water demand helps in future water demand predictions, optimization of the water system operations and models the water allocation system [4].

The objective of this paper is to develop a prediction model which helps in anticipating the future water demand which is mostly hooked on uncertain parameters. This paper discusses about Support Vector Machine (SVM) used for forecasting water demand which helps in optimal and effective water demand management [5]. The paper is structured as follows. Section 2 deliberates on the research works that has used prediction methods on organization of water resources. The brief overview of the methodology used is explained in section 3. Section 4 elaborates the dataset used in the study. Finally section 5 and section 6 contains the outcomes and the conclusions.

II. RELATED WORK

Data mining techniques helps in extracting hidden intelligent data from a huge collection of data. Different water demand predicting techniques have been suggested with various features linked to precise goals, data availability, forecast horizons, and used variables. To analyse the performance in forecasting water demand, prediction techniques like artificial neural network, support vector machine, deep neural network, Gaussian process regression, multiple regression and random forest can be used. This section provides an outline of the different methods used in predicting water demand.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Sarakutty T. K.*, Research Scholar, Rayalaseema University, Kurnool, India, E-mail: sarajobythomas@yahoo.co.in
Ravikumar K., Research Scholar, Kalinga Institute of Industrial Management, Bhubaneshwar, India, , E-mail:ravinaac@gmaill.com
Hanumanthappa M., Department of Computer Science and Applications, Bangalore University, Bengaluru, India,
E-mail: hanu6572@hotmail.com

Predictive modeling technique like Regression analysis was used in water demand prediction due to its simplicity and reliability. Artificial neural networks which is a self-adaptive model that captures nonlinearities in data,
incoherence of data and polynomial components are extensively used in predicting water demand. Fuzzy inference systems and Neuro-Fuzzy systems are also used in predicting water demand [6]. Some of the optimization methods used in water demand prediction are Particle Swarm Optimization and teaching-learning-based optimization [7, 8].

Adamowski used multiple linear regression, time series analysis and artificial neural network in order to forecast summer peak consumption [9]. Ghiassi et al used artificial neural network model in order to forecast the demand for water and compared it with artificial neural network model and auto regressive integrated moving average model [10]. Arandia et al. compared actual amount of water produced with the forecasted one using self-correcting seasonal moving average [11]. Candelieri used two stage support vector machine, one stage for clustering and the second stage for short term prediction of water demand [12]. Brentan et al. used a hybrid technique by joining support vector machines with Fourier time series for urban water demand models [13]. Brentan et al. used combinations of self-organizing maps and RFs [14, 15].

III. METHODOLOGY

Support vector machine (SVM) originated in statistical learning theory and uses classification and regression methods. The SVM model is represented in "Fig 1".SVMs focus on nonlinear separable data problems. To accomplish this detachment, a SVM discovers the perfect hyper plane that makes best use of the separation among two groups, in this way limiting the margin error. Here the input data is denoted in a n-dimensional plane which is not divisible linearly and is plotted on a space in bigger dimension so as to acquire linear regression [15, 16].

When SVMs are applied to regression problems it is called as (SVR) Support Vector Regression. Where x, the input space is mapped to Φ(x), high dimensional feature space in a non-linear fashion. Let (x1,y1), (x2,y2), ... (xm,ym) be a training data set with one input and one output. Then the aim is to calculate a kernel function, f(x) which maps the inputs to the outputs of the training set. Support Vector regression finds this function which has at utmost deviance as of the real training outputs yi.

The least difficult technique to stretch out the support vector regression on to data which is nonlinear is pre-processing the training dataset by utilizing a mapping function ϕ from the input space to the feature space. The significant outcome is that, instead of expressly mapping every data into the new space a kernel function can be used. The kernel function empowers activities to be executed in the input space instead of the feature space. Radial basis functions, Polynomial, Sigmoid are some of the kernel functions.

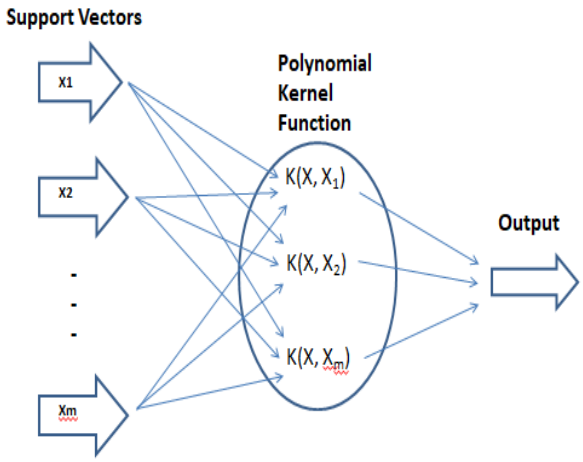

Fig 1: SVM Model

Toward the finish of the clustering methodology, a constrained cluster set is recognized; each one is viewed as a data set in the second arrangement phase of the projected methodology. The primary p segments of every vector of the cluster forms the input variables for the SVM regression model and compares the water utilization for various months. The objective variable to foresee is the hth section of the first dataset, were h varies from p+1 to 24. SVM regression model is prepared on behalf of each h, and every one gathers a pool of SVMs for the particular group.

To utilize the learned models for prediction, the particular SVM pool is recovered monthly from the first stage cluster. At the point when another grouping of p month to month request esteems is accessible, it is utilized as contribution to the chosen SVM pool, and each SVM regression model in the pool gives the anticipated value to the related month.

IV. DATASET

The data set used in the study contains consumption of water data obtained from BWSSB billing database and water cess records; previous water demands and the estimated population size. Total water consumption data, was considered and domestic, industrial and commercial sectors were not identified separately. The water demand data is complex in nature due to uncertain nature of the data. Per capita demand varies in normal conditions and in drought conditions. The dependency of water is dependent on the various features like geographic location,population, climate, size, and economic condition of community, metered water supply, extent of industrialization, water cost and supply pressure. Water demand also fluctuates seasonally, daily and hourly. The water demand database consists of the average value of the daily water consumption for different years, the total water supply connections for each year, actual water consumptions month wise and the total connections. For the month to month analysis, the data was aggregated into month to month sums, normalizing every aggregate for the month duration with the goal that all information accept a 30 day in a month. From the water bill the amount of water delivered for a particular month is available.

Usually for each ward the water is supplied on alternate days and the date and amount of water supplied ward wise is also available. Therefore, from the water delivery statement it is only possible to obtain information on the amount of water delivered on any particular date. The number of occupants in a particular building is also inexact therefore from this data it is not usually possible to get the exact amount of water used by a household. Therefore to achieve a precise association between the class and non-class attribute daily water usage is required. The population prediction values and the number of connections are used to predict water demand from the base year. Equal water distribution based pre-processing approach was used for estimating the daily water usage. Here average water usage per day is calculated by dividing the amount of water supplied by the number of days among two consecutive supplies [17].

Pre-processing was done to transform the raw data. From the date time variable, time was mined and allocated a categorical value of one to twenty four. Month values was mined and allocated a categorical value 1to 12. Continuous feature variables were rescaled to the range of o and 1 by finding the variables maximum value and minimum value.Let x be the original data point, then the normalized data point, x'was computed by subtracting the minimum value of x from x and dividing it by maximum value of x minus minimum value of x.

This rescaling was done to evade assigning more weight to features with higher values [4, 18]. The family unit size (for example normal people per family unit) is the explanatory variable chosen. Based on the number of family individuals per capita water utilization changes. Be that as it may, the variation may be negative or positive, based on whether water utilization increment pretty much relative to the expansion in family unit size [19]. Chronicled information like water utilization, population growth, population number, number of guests, and people per house were gathered from the relevant sources. An old style path for imparting vulnerability to explanatory variable is by assigning probability distribution function. Other than the choice of the probability distribution function, there is the errand of picking reasonable factors that portray the probability function. These incorporate, for example, parameters such as standard deviation, mean, least or most extreme permitted values. The family size can't be anticipated with exactness since it fluctuates radically. The family unit size, normal people per family, was expected to pursue a typical appropriation, as proposed by statistic information which additionally directs a mean value of 0.02 and standard deviation of ten percentage of the mean [3].

The predictions was validated by one month ahead, short term and k months ahead, long term estimates. As a result of the partial dependency of the data, arbitrary subset of data cannot be chosen for training and validation, however partial ordering of data should be taken care inside training and validation sub sets to catch temporal elements. For estimates of smaller duration, the training data set comprised of previous perception selected before the prediction period which was the test subset. This was rehashed multiple times throughout the previous year to mirror a true situation whenever data is constantly added to the model. The tailored model is refreshed and used to figure request in the following month. The evaluation measurements were averaged every year to give a total proportion of one month prediction. For long time prediction, the test data set enclosed all perceptions throughout the previous year in the data set. Predictions were made for every one year in the test data set by maintaining static training data without including extra information.

Again appraisal measurements were averaged for every one year anticipated. For small and long interval appraisals, the test datasets incorporates all family units in the examination. That means no subclass of water utilization is perceived within the month being anticipated, to line up with genuine prediction models. For feature-based machine learning models, the future predictor values were assumed to be known when making predictions.

V. RESULT AND DISCUSSION

The outcomes of the study suggests that usage of water is emphatically transiently related, and that representing the temporary dependence is successful at delivering precise predictions than demonstrating relationship with exogenous features. The water demand patterns were identified for five months through two level clustering procedures where seven different types of users were selected. For each cluster one SVM is trained for different months with input feature as monthly consumption, p=5 and target variable as water consumption.

The SVM Parameters are adjusted independently by minimizing RMSE (Root mean square root error) which processes the average prediction error for present data. RMSE is calculated using (1), where xi is the measured water demand and xi ' is the projected water demand at time i [15].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i - x_i'} \quad (1)$$

Mean absolute error (MAE) calculated using (2) is used to measure the forecasting error to detect anomalies like deviations from real consumption pattern to predicted forecast. Deviances from projected consumption manners can happen due to meter faults, water theft or false transmission data.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - x_i'}{x_i}\right| \quad (2)$$

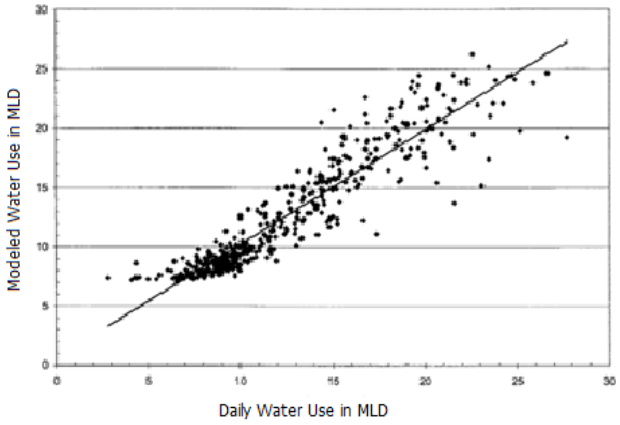

Fig 2: Comparison of the daily water use pattern

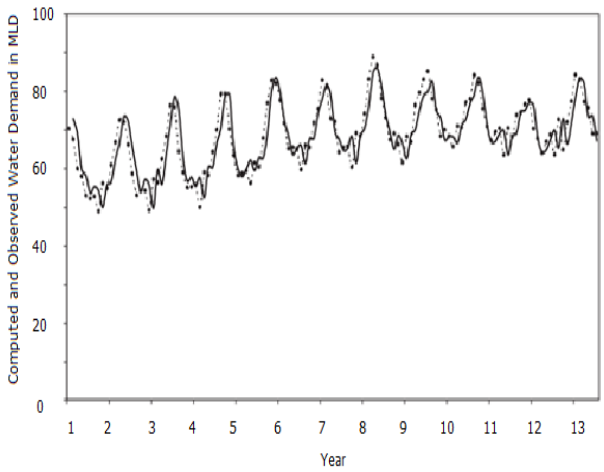

Fig 3: Computed and Observed Water Demand

The Mean Absolute error obtained by each SVM pool on each cluster is calculated. MAE values may vary based on the type of customer. In domestic case, MAE does not indicate much variability and therefore only a few set of real regular patterns are considered inconsistent with respect to prediction.

The daily average demand is predicted for every month of the predicting years. The monthly forecast assessment averaged over 12 months between modeled daily water use and recorded daily water use is represented in "Fig 2". Radial basis function with 35 bounded support vector was used for this. The long term water conservations resulted in approximately 24% in water consumption. Generally, the forecasting from the model are rational, where in the standard deviation is of 0.85. The model is evaluated using different metrics like RMSE, R-Squared and MAE. The nearness of data to the predicted value is measured using R-squared which was 8.03. It indicates the percentage variation of response variable. The RMSE and MAE values were 8.03 and 5.87 respectively. The prediction for dataset is reiterated for different time interval. The R-square value is helpful in analyzing the performance of the methods for different time durations. The performance of all the methods surges significantly if the time duration is large. As the duration is expanded the unexpected deviation in the data is diminished extensively, which has caused an increase in the performance. Computed and observed water demand obtained for thirteen years is represented in "Fig. 3 ".

VI. CONCLUSION

Expanded checking in water distribution frameworks permits the use of data mining strategies all the more likely comprehend the framework and gauge future pressure driven states. Advancement calculations help to discover great models for different prediction procedures. The consequences of our examination propose that usage of water is unequivocally transiently related, and representing temporal dependence is further viable at delivering precise figures than displaying relationship through exogenous highlights. The framework is utilized to dissect the water utilization design in order to foresee the future water request in the investigation region. Despite the fact that the outcomes got from this examination are good in determining water request, a confinement would be the utilization of low number of influential qualities in the dataset. This can be additionally improved by including more characteristics having high influence on water utilization. The examination was centered uniquely around local interest, which is most questionable part for the future water request. Surveying the water request influence because of environmental change was not considered because of the month to month time step. The assessment of the best model structure helps in reducing the quantity of sources of information,lessens the calculation of computational time and the problems at the extrapolation limits. Improved information assortment and strategies are important to expand expectation precision for family level month to month water use. Higher-goals covariates may conceivably build conjecture quality, particularly long haul estimates. Conjectures may likewise be improved by utilizing utilities-explicit indicators, estimating, discount projects and water limitation approaches, and their associations with extra family level statistic data.

REFERENCES

1. Mohamed M. Mohamed; Aysha A. Al-Mualla , Water demand forecasting in Umm Al-Quwain using the constant rate model, Elsevier BV, Desalination, ISSN: 0011-9164, Vol: 259, Issue: 1, Page: 161-168, 2010.
2. Isaac Duerr, Hunter R.Merril, ChuanWang, RayBai, Mackenzie Boyer, Michael D.Dukes. NikolayBliznyuk, Forecasting urban household water demand with statistical and machine learning methods using large space-time data: A Comparative study, Environmental Modelling & Software, Volume 102, April 2018, Pages 29-38.
3. Ibrahim Almutaz, AbdelhamidAjbar, Yasir Khalid, Emad Ali , A probabilistic forecast of water demand for a tourist and desalination dependent city: Case of Mecca, Saudi Arabia, www.elsevier.com/locate/desal , Desalination 294 (2012) 53–59.
4. Vijai,P. and Sivakumar,P. B. (2016) "Design of IoT Systems and Analytics in the Context of Smart City Initiatives in India." *Procedia Computer Science* 92:583-588.
5. Praveen Vijai,BagavathiSivakumar P, Performance comparison of techniques for water demand forecasting, 8th International Conference on Advances in Computing and Communication (ICACC-2018) , www.elsevier.com.
6. Mohsen Nasseri , Ali Moeini, MassoudTabesh, Forecasting monthly urban water demand using Extended Kalman Filter and Genetic Programming, Expert Systems with Applications 38(2011) 7387-7395
7. Zhu,X. and Chen,J.(2013) "Urban water consumption forecast based on PQPSO-LSSVM" Proceedings - International Conference on Natural Computation: 834-837.
8. Ji,G. and Wang,J. and Ge,Y. and Liu,H. (2014) "Urban water demand forecasting by LS-SVM with tuning based on elitist teaching-learningbased optimization" Chinese Control and Decision Conference, (CCDC): 3997-4002.
9. J. F.Adamowski, "Peak daily waterdemand forecast modeling using artificial neural networks," Journal of Water Resources Planning and Management,vol.134,no.2,pp.119–128,2008.
10. G.Ghiassi, D.K.Zimbra and H.Saidane, "Urban water demand forecasting with a dynamic artificial neural network model," Journal of Water Resources Planning and Management, vol.134, no.2, pp.138–146, 2008.
11. B.Eck, E.Arandia, A.Ba and S.McKenna, "Tailoring seasonal time series models to forecast short-term water demand," Journal of Water Resources Planning and Management, vol.142, no.3, 2016.
12. A. Candelieri, "Clustering and support vector regression for water demand forecasting and anomaly detection," Water (Switzerland), vol.9, no.224, 2017.

13. M. Herrera, E. Luvizotto Jr., B. M. Brentan, G. Meirelles, and J. Izquierdo, "Correlation Analysis of Water Demand and Predictive Variables for Short-Term Forecasting Models," Mathematical Problems in Engineering, vol. 2017, Article ID 6343625,10 pages, 2017.
14. M. Herrera, J. Izquierdo, B. M. Brentan, E. Luvizotto Jr., and R. P´erez-Garc´ıa, "Hybrid regression model for near real-time urban water demand forecasting," Journal of Computational and Applied Mathematics, vol.309, pp.532–541, 2017.
15. Julia K. Ambrosio, Bruno M. Brentan , Manuel Herrera , EdevarLuvizotto Jr., LubienskaRibeiro, and Joaqu-n Izquierdo, Committee Machines for Hourly Water Demand Forecasting in Water Supply Systems, Mathematical Problems in Engineering Volume 2019, Article ID 9765468, 11 pages.
16. Carlos Peña-Guzmán, JoaquínMelgarejo and Daniel Prats, Forecasting Water Demand in Residential, Commercial, and Industrial Zones in Bogotá, Colombia, Using Least-Squares Support Vector Machines, Mathematical Problems in Engineering, Volume 2016.
17. Mahmood A Khan ,MdZahidul Islam , MohsinHafeez , Irrigation Water Requirement Prediction through Various Data Mining Techniques Applied on a Carefully Pre-processed Dataset Journal of Research and Practice in Information Technology, Vol. 43, No. 22, May 2011.
18. Ishmael S. Msiza, Fulufhelo V. Nelwamondo, TshilidziMarwala, Water Demand Prediction using Artificial Neural Networks and Support Vector Regression, JOURNAL OF COMPUTERS, VOL. 3, NO. 11, NOVEMBER 2008.
19. J. Schleicha, T. Hillenbrand, Determinants of residential water demand in Germany, Ecol. Econ. 68 (2009) 1756–1769.

AUTHORS PROFILE

Sarakutty T. K. is a research scholar in the department of computer science, RayalaseemaUnversity where she furthers her research on optimal allocation of water resources using data mining techniques. She has authored more than five research papers in international journals and has presented her research works in international conferences. Her areas of interest and research are Data Mining, Optimization Techniques, Predictive Analytics and Algorithms.

Dr. Hanumanthappa M. is a Professor in the department of Computer Science and Applications, Bangalore University, Bangalore. He received his doctorate in Computer Science from Bangalore University. He has published more than 100 research papers and articles in international journals and conference proceedings. His research areas include Data Mining, Machine Learning, Security and Natural Language Processing.