Biological Forum – An International Journal **7**(1): 1469-1473(2015)

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

Evaluation of a Data Mining model in Predicting of "Average Temperature" and Potential Evapotranspiration Month for the next Month in the Synoptic Weather Station Yazd

Seyyed Hassan Mirhashemi* and Mehdi Panahi**

**Ph.D. student of Irrigation and Drainage, Water Engineering Department, University of Zabol, IRAN*
***Assistant Professor of Water Engineering Department, College of Agriculture, University of Zanjan, IRAN*

(Corresponding author: Seyyed Hassan Mirhashemi)
(Received 12 April, 2015, Accepted 14 May, 2015)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: In this paper have been evaluated the ability of M5P Tree models to estimation "average temperature" and "Monthly potential evapotranspiration months later," for Yazd synoptic weather stations. The data used in this article, are the average monthly data of Yazd weather station, including: "average temperature", "sunshine hours", "dew point", "relative humidity", "average wind speed" and "saturation vapor pressure deficit" in forty-six-year period from 1960 to 2005 AD. Output variables used, are "average temperature" and "Monthly potential evapotranspiration months later," as monthly basis. After introducing the weather data as mean monthly to the algorithm as input variables and monthly potential evapotranspiration months later, as the output variables, to the M5P algorithm were evaluated using "correlation coefficient", "Root Mean Square Error" and "mean absolute error". According to the three statistical indexes, M5P Tree model have better function in estimating the monthly average temperature for the months later.

Keywords: Data mining, M5P Tree model, Penman-Monteith equation, average temperature, synoptic weather station, Yazd

INTRODUCTION

Meteorological data those are measured and archives at different stations included a large volume of information and are increasing their volume over time. Accordingly, most felt the need for new methods of "data mining" of them. In some cases, a lot of variables are used that may be some of them are not measure in all weather stations. Therefore is necessary to make use of modern techniques such as data mining.

Definitions of data - mining. Data mining has many broad definitions. The definitions lot depends on the individual backgrounds and points of view. So we can say that the data mining is a set of methods in process of knowledge discovery that used to recognize patterns and undisclosed relationships in data. Data mining can also be said is a process recognition valid, new, useful and understandable pattern, from data. Data mining is a technique that combines hypothesis tests and derives data- discovered. In the Assuming tests, researchers can test ideas against the data to confirm or refute its validity. Vandenberg and colleagues (1999) explain that discovery; the researcher draws conclusions from the data and allows the data to accept the result. The most data mining problems is solved using a combination of both methods. For example, the result may be a new hypothesis that can be tested and the test will be approved or rejected. Data mining is the process of selecting, identifying and modeling large amounts of data (Giudici, 2003).

In another definition, the process of selecting, exploring and modeling large data mining officials. To discover hidden relationships and achieving results clearly beneficial for the owner of the database (Meshkani, 2009). Data mining is a process that uses various tools to analyze data, to the physical changing patterns and relationships found in different data sets. The main difference between data mining and statistics, that is data mining is one approach without the default. While most conventional statistical techniques are needed to default. And statistical professionals are searching equations to match the defaults. In contrast, data mining algorithms can automatically develop these equations from information contained in the data set (Cabena *et al.*, 1998).

M5P algorithm. Regression tree models with tree concepts are generally constant in leaves (Witten and Frank, 2005). They simulate the regression function piece - wise (and therefore are nonlinear).

M5P a binary regression tree models in their final nodes (leaves) are linear regression functions that can produce continuous numerical attributes. Models based on tree splitting method and conclusions are made.

Generates a tree model requires two steps. The first step involves the use of a divergence metric to produce a decision tree. Branching criterion for M5P model tree algorithm is the behaved class values that reach a node as the quantification of the error and the expected reduction in error as a result of testing each attribute at that node is calculated (Quinlan, 1992).

The formula for calculating the standard deviation reduction (SDR) is as follows:

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

In which:

T = number of samples indicate that the node is;

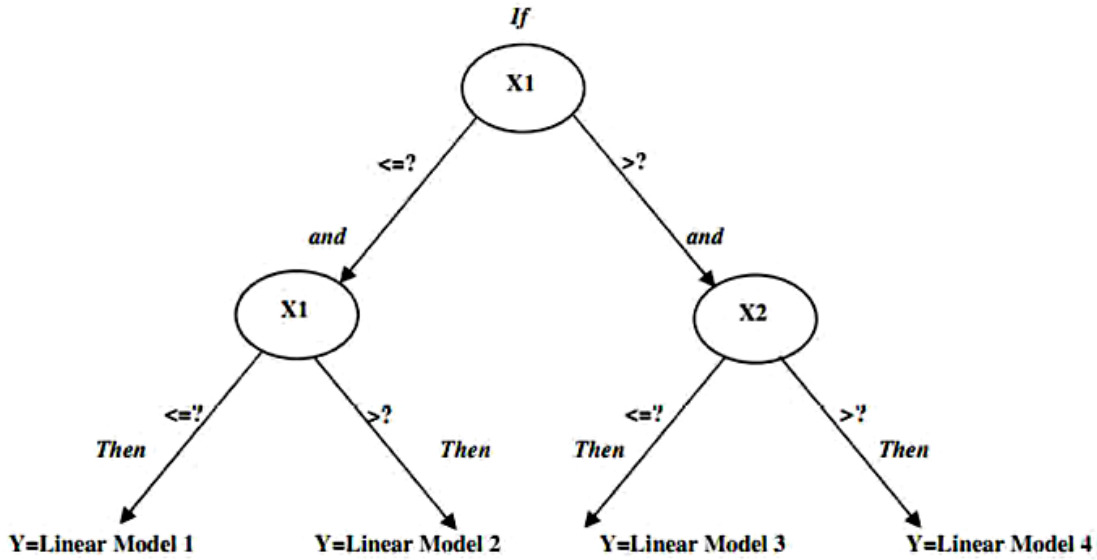
T_i = number of samples that represent the ith test have the potential to rise;

sd = standard deviation is indicated.

Because of the split process, data in child nodes have lesser than standard deviation from the parent node and hence Purer. After maximize all possible ramifications, M5P select which attribute will reduce the expected maximum.

This division often creates a tree-like structure that is over-fitting. To overcome the problem of over-fitting, the tree should be pruned back, For example, by replacing a sub-tree with a leaf.

Therefore, the second step of the design model including tree pruning, tree removal and replacement of trees, plants grown with linear regression functions. This technique generates a tree model, the parameter space into regions (subside) and split each of them makes a linear regression model. For more details on the model tree M5P refer to (Quinlan, 1992).



MATERIALS AND METHODS

Yazd province is located in the center of Iran and in zone of central mountain of Iran in area at 29.48 to 33.30 north latitude, and 52.45 to 56.30 east longitudes. Yazd province limited from north and west to Isfahan province, from northeast to Khorasan province, from southwest to Fars province and from southeast to Kerman province. Yazd province area is about 72156 km2, and is about 4.37 percent of Iran area. Weather of Yazd province due to being on a dry zone of the world has cold and relatively wet winters and hot and long and dry summers. Under the general population and housing census in 1375, Yazd provinces population was 750,769 people, which have formed 15.75% urban population and 85.24% rural population. The word “Yazd” means pure and holy and “Yazd city” is also means “City of God” and “Holy Land”. For reason climatic conditions of Yazd province, the agricultural situation is not desirable., and the possibility of exploiting the surface water in agriculture is very low. Specific conditions of desert edge like low rainfall, moving sands, the phenomenon of desertification, poor pastures, lack of water resources has led to 28% of the area of Yazd hasn't productivity economic.

The agricultural areas in Yazd, is including plains of Yazd, Ardakan, Bahadoran, Bahabad, Herat, Marot, Chahak and Abarkooh. Agricultural products, is including pomegranate, pistachio, almonds, beans, sunflowers, grapes, cotton, sugar beet and sesame. Yazd province in terms of geology is having huge mineral reserves. Major mines in the province as "Iron Choghart", "Marble Bouraghi", "sandstone Matkasaneh", " kootekplumb" were having important and constructive role in economic transformation and development of province.

The data used in this article, are the average monthly data of Yazd weather station, including: "average temperature", "sunshine hours", "dew point", "relative humidity", "average wind speed" and "saturation vapor pressure deficit" in forty-six-year period from 1960 to 2005 AD. The 75 percent of data used as model generation with using M5P Tree model, and 25 percent of data used as test model. To predicting average temperature and percent of relative humidity as monthly for next month is selecting sex variables including sunny hours (h), dew point temperature (c), average relative humidity (%), average wind speed (m/s), saturation vapor pressure deficit (mbar) and average temperature as monthly before month as input variables and average temperature and average relative humidity as output variables.

RESULTS

To calculate "average temperature" and "monthly potential evapotranspiration next month" was used average monthly data series of Yazd station. Values "average temperature" and "monthly potential evapotranspiration next month" estimated from the M5P Tree model were compared with "average temperature" and monthly potential evapotranspiration next month", calculated by synoptic station by "correlation coefficient", "Root Mean Square Error" and "mean absolute error".

As can be seen in Table 1 M5P Tree model with a correlation coefficient of 0.9772, RMSE of 0.6825MAE of 0.4971 for estimation monthly average temperature and a correlation coefficient of 0.8642, RMSE of 0.8337 MAE of 0.6712 for estimation monthly potential evapotranspiration next month was estimated. From this estimation resulting that M5P Tree model to estimate the average temperature has better estimation to estimate the Monthly potential evapotranspiration next months".

Table 1: Comparison correlation coefficient, Root Mean Square Error and mean absolute error to estimate the average temperature and Monthly potential evapotranspiration next months by M5P Tree model.

Method	R^2	MAE	RMASE
average temperature	0.9772	0.4971	0.6825
Monthly potential evapotranspiration next months	0.8642	0.6712	0.8337

As showed in Table 1, the M5P Tree model in estimation monthly average temperature for next month according to correlation coefficient, Root Mean Square Error and mean absolute error had better function to estimate the Monthly potential evapotranspiration next months. An advantage of Tree models such as M5P is its access to complex many simple linear functions that can use to predicting potential evapotranspiration next months (Fig. 1).

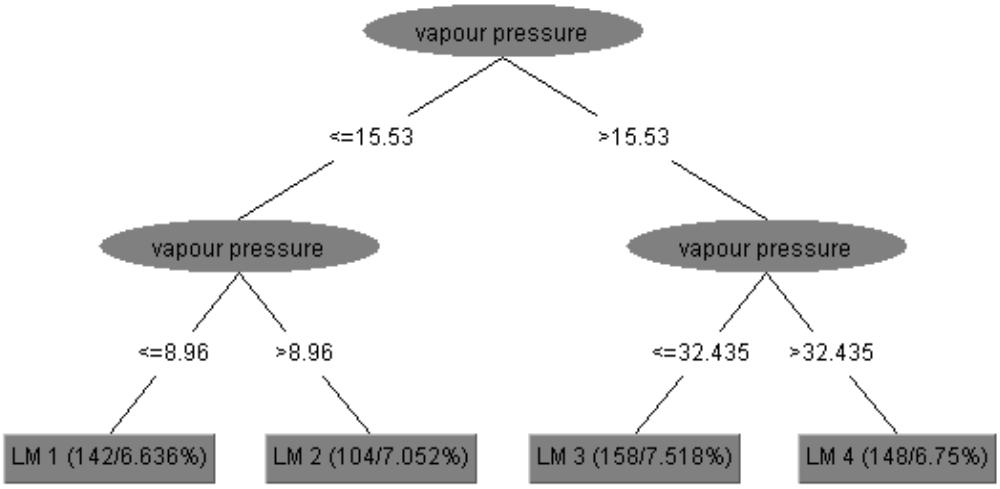


Fig. 1. The tree diagram from performance of M5P models to predicting of monthly average temperature for the next month.

From Fig. 1 Resulted that first important parameter to making M5P is monthly saturation vapor pressure deficit that in first was devising tree to half e ≤15.53 and e > 15.53. Thenceforth, to according to divergence scale in both stem again monthly saturation vapor pressure deficit has more important. The divergence continued in this procedure that in the end obtained four equation that each applies toparticular section.

LM num: 1
avreg = -0.1212 * Humidity - 1.4551 * wind + 0.6734 * dewpoint + 0.8066 * vapour pressure + 0.0332 * targe + 12.1585
LM num: 2
avreg = -0.03 * Humidity - 1.9453 * wind + 0.3604 * dewpoint + 0.804 * vapour pressure + 0.2314 * targe + 7.1522
LM num: 3
avreg = -0.0242 * Humidity - 1.7595 * wind + 0.0063 * sun + 0.2453 * dewpoint + 0.4961 * vapour pressure + 0.241 * targe + 11.6279
LM num: 4
avreg = 0.1241 * Humidity - 1.4566 * wind + 0.0067 * sun + 0.0593 * dewpoint + 0.4123 * vapour pressure + 0.0368 *targe + 12.53

A. Sensitivity analysis

To determine the most important factor for modeling monthly average temperature for the next month via M5P Tree model were compared by changing the input data and using the statistical parameters, which contains the correlation coefficient, Root Mean Square Error and mean absolute error. When compared first row were include five meteorological parameters have most regression and less square root error and mean absolute error. As a result, five parameters were used have the greatest impact in the function M5P Tree model. Then the four parameters are located in the second row and are including the parameters of average monthly humidity, sun hours, wind speed and dew point, the four parameters are located in five row are including parameters of average monthly humidity, sun hours, saturation vapor pressure deficit and dew point in the second and third ranks respectively in the positive impact M5P Tree model to proper functioning in the estimating average monthly evapotranspiration for the next month.

Table 2: The analysis sensitivity of M5P Tree model to estimate monthly average temperature for the next month.

Combination input parameters	R	MAE	RMSE
*n,w,RH,dwe,e	0.997	0.505	0.701
n,w,RH,dwe	0.994	0.707	0.962
n,w,RH,e	0.988	1.074	1.360
n,w,dwe,e	0.986	1.206	1.496
n,RH,dwe,e	0.994	0.720	0.968
n,w,dwe	0.933	2.520	3.281
n,w,e	0.985	1.230	1.528
n,w,RH	0.930	2.592	3.340
n,w	0.896	3.071	4.056
n,RH	0.928	2.641	3.398
w,dwe	0.695	5.426	6.566
n,dwe	0.933	2.513	3.284

RH: average relative humidity (percent), T: average daily temperature '(C), W: Wind speed (m/s), e: saturation vapor pressure deficit (mbar), DEW: dew point (c), n: Sunny Hours (h).

CONCLUSION

From this study it can be concluded that: Techniques of "data mining" such M5P Tree model can be used to estimate average evapotranspiration and monthly average temperature of the next month. M5P Tree model with an estimate of monthly average temperature for the next month is shown that can have a high capacity to estimating meteorological parameters. This model can be used in a variety of stations that are deficient in recorded meteorological parameters. It was concluded from Table 2.

Sensitivity to M5P Tree model with enter the six parameters, including average temperature (c), sunny hours (h), dew point temperature (c), average relative humidity (%), and saturation vapor pressure deficit (mbar) as input variables have the best performance, relative to the composition of the other parameters in Table 2.

REFERENCES

Meshkani, A. Nazemi, A.R. Introduction to "data mining", Neishabour branch of Islamic Azad University Press, 1388, pp. 456.

Cabena, PH. Stadler R., Verhees J., and Zanasi. (1998). Discovering data mining: From concept to implementation, IMB, New Jersey, 195 pp.

Giudici, P. (2003). Applied data Mining: statistical methods for business and industry. Wily, London. pp. 364.

Quinlan,JR. (1992). Learning with continuous classes. Proceeding of Australian Joint Conference on Artificial Intelligence. World Scientific Press: Singapore; 343-348.

T Crows Corporation, (1999). Introduction to data mining and knowledge discovery, third ed., Postmac, MD. Available at: www.twocrows.com, (April 29, 2000).

Vanderberg, H. Sogard P. and Motoroni S. (1999). MineSetTM 3.0 Enterprise Edition Tutorial for Windows, Doc. No. 007-4006-001, Silicon Graph.