International Journal of Agriculture and Crop Sciences.
Available online at www.ijagcs.com
IJACS/2015/8-1/5-8
ISSN 2227-670X ©2015 IJACS Journal

IJACS

Evaluation ten models of weka software to predict the potential evapotranspiration month for the next month in the synoptic weather station Arak

Seyyed Hassan Mirhashemi[*1], Mahmod Tabatabayi[2]

1. PhD student of Irrigation and Drainage, Water Engineering Department, Zabol University
2. Assistant Professor, Faculty of water and soil, Water Engineering Department, Zabol University, Iran

Corresponding Author email: hassan.mirhashemi@yahoo.com

ABSTRACT: In this paper have been evaluated the ability of ten models of weka software to estimation "Monthly potential evapotranspiration months later,"for Arak synoptic weather stations. That is including additiveRegression, Bagging, LinearRegression, Zero, M5P, Kstar, M5Rules and REPTree. The data used in this article, are the average monthly data of Arak 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, was "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, models "data mining" were evaluated using "correlation coefficient", "Root Mean Square Error" and "mean absolute error". According to the statistical indexes, Tree Bagging models have better function in estimating the monthly average temperature for the month.
Keywords: "Data mining" weka, Penman-Monteith equation, potential evapotranspiration, synoptic weather station, Arak

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 [PH Cabena., Stadler R., Verhees J. and Zanasi .: 1998]. 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 [Giudici, 2003]. Data mining is the process of selecting, identifying and modeling large amounts of data. 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 [Mamashkani, 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 [Two Crows , 1999]. 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].

MATERIALS AND METHODS

The study area

To conduct this research, the synoptic weather station ID of climate data were obtained from the Meteorological Branch of the organization.

Characteristics of the study area

Arak**:** Arak city limited from north to Tafresh, from west, to Hamadan and Malayer from east to Mohalat city and from south to the city of Khomein and shazand. City has an average height of about 1,700 meters above sea level, and in terms of area, the second largest city of the province after the city of Saveh. Arak city in terms of its climatic factors (such as proximity to Desert Mighan, altitude, etc) climate fluctuations. Generally fairly mild summers and cold winters and relatively cool climate characteristics of Arak city. The center of Markazi Province is Arak, is located near the center of Iran and the circuits 33 degrees 30 minutes to 35 degrees 35 minutes north latitude and 48 degrees 57 minutes to 51 degrees east of the Greenwich meridian. In various parts of Markazi Province, diverse semi-arid climates, temperate and cold mountains there are mountains.

Procedures for estimating potential evapotranspiration

Penman-Monteith method as one of the most reliable methods for estimating ETo is used by professionals. In this method, the reference plant is supposed grass cover that it's height is 12 cm and "Albido factor" is 23 percent (in real grass plant is 25 percent).

Penman-Monteith equation, equation (1) that was used to estimate evapotranspiration in this paper is as follows:

$$ET_O = \frac{0.408[(R_n - G) + \gamma \frac{900}{T+273} U_2.(e_s - e_a)]}{\Delta + \gamma(1 + 0.34U_2)} \qquad \text{Equation (1)}$$

In this regard,

ETo = The standard evapotranspiration of reference grass depending on the mm per day,

es, ea = vapor pressure and the actual pressure of water vapor in the air, according Millibar,

U2 = wind speed per day at a height of two feet on the ground m/s ،G ،Rn MJm-2d-1

Δ = in terms of slope of the curve the saturation vapor pressure (es) over the temperature range T)),

γ = constant Sicometry Kpa c-1

3-5 - statistical indicators used and Comparison in the algorithms

Regression equation (2) and mean absolute error of equation (3) for the comparison between Penman-Monteith equation and the six algorithms "data mining" is used as following:

$$1 - (\sum_{i=1}^{n}(Yobs - Ypred)\text{^}2 / \sum_{i}^{} \quad \text{Equation (2)} \quad n)\text{^}2) = R^2$$

$$\left| \frac{1}{n} \right. \text{MAE} = \quad \text{Equation (3)}$$

N: number of observed data,

yobs: potential evapotranspiration observed

ypred: potential evapotranspiration predicted

ymean: potential evapotranspiration average.

To choose the most suitable algorithm "data mining" in the estimate " potential evapotranspiration next month", after calculate potential evapotranspiration by "Penman-Monteith" equation as output data, and the average monthly weather data as input data for a period of forty-six years for the software "data mining" took place, And then with compare regression coefficient and mean absolute error between that is done by the software "estimates derived" from "Data Mining Algorithms" and the answer the "Penman-Monteith equation" by selecting the highest correlation and the lowest average absolute error between them, we obtained most suitable algorithm for the station.

Many scientists have studied the Penman-Monteith equation to estimate ETO (Allen et al, 1998; De uder et al 1995). Jensen (1990) were analyzed the performance of 20 different methods against "ET" was measured for 11 stations located in different climatic zones of the world. Penman-Monteith method was ranked as the best method for all climatic conditions. Application of Penman-Monteith equation FAO - 56 needs the data of "sunlight", "Wind speed", "Temperature", "vapor pressure" and humidity, but all these input variables are not readily available in every location. In developing countries, the correct data collection problems and these can all climatic variables encountered in application Penman-Monteith equation FAO - 56 is considered a serious problem. Software Weka has developed in Waikato University in New Zealand and its name has been extracted from the phrase "Waikato Environment for knowledge Analysis".

Also Weka, called founder wild bird that does not fly and is found in New Zealand. The system is written in Java and has been published based on sweeping the GNU General Public License. Weka almost runs on any platform and is tested under Linux, Windows, Mac, and even a digital receptionist person. This Software is sweeping an interface to many different learning algorithms, that through it's the procedure pre- process, post - process and evaluate learning schemes on all data sets, are applicable. This environment is includes issues procedures for all standard "data mining" such as regression, classification, clustering, "exploring the association rules" and "feature selection".

Considering that the data are an integral part many of the tools - data processing and visualization has been provided. All algorithms get their entry into as relational table to ARFF format. This format of the data can be generated from a read file or from a database by a query.

In this study, we was used eight different models of Weka software which was include additiveRegression, Bagging, LinearRegression, Zero, M5P, Kstar, M5Rules, REPTree to predict "potential evapotranspiration months later".

Also used monthly meteorological data of "Arak synoptic weather stations" as "inputs data" that includes: "The average temperature» (c), "Sunny Hours" (h), "Dew point» (c), "Relative humidity" (percent), "Average wind speed" (meters per second) and Saturation vapor pressure deficit (mbar) in forty-six-year period from 1960 to 2005. 75% of the data as a production model using Bagging model and 25% of them was used as the test model. To predict monthly potential evapotranspiration next month used six variables: "Sunny Hours" (h), "Dew point» (c), "Relative humidity" (percent), "Average wind speed" (meters per second) and Saturation vapor pressure deficit (mbar) and "Average temperature" That was considered all as monthly basis month after month as the input data and the monthly potential evapotranspiration next month as the output data.

RESULTS

To calculate "monthly potential evapotranspiration next month" was used average monthly data series of Arak station.

Values "monthly potential evapotranspiration next month" estimated from the eight models were compared with monthly potential evapotranspiration next month", calculated by Penman-Monteith equation by "correlation coefficient", "Root Mean Square Error" and "mean absolute error". As can be seen in Table 1 Bagging tree model with a correlation coefficient of 0.8369, MAE of 0.6579 and RMSE of 0.8467 the appropriate model was selected to estimate the "Monthly potential evapotranspiration next months".

Table 1 . Comparison ten models weka software with three statistical indices

Statistical indices models	R	MAE	RMSE
additiveRegression	0.7997	0.7436	0.9319
Bagging	0.8369	0.6579	0.8467
Kstar	0.8219	0.6294	0.886
LinearRegression	0.142	1.1608	8.8889
LWL	0.7062	0.9234	1.0968
M5P	0.1736	1.0132	7.3521
M5Rules	0.1736	1.0132	7.3521
REPTree	0.7969	0.6703	0.9473
Vote	-0.1142	1.3895	1.5484
Zero	-0.1142	1.3895	1.5484

Table 2. The combination of input parameters to estimate the monthly potential evapotranspiration for the next month, using Bagging model.

Statistical indices combination of input parameters	R	MAE	RMSE
T,n,w,RH,dwe,e	0.8467	0.6579	0.8369
T,n,w,RH,e	0.8382	0.6392	0.8434
T,n,RH,dwe,e	0.9502	0.3463	0.4825
T,n,w,e	0.8304	0.663	0.8617
T,n,w,RH	0.8379	0.6327	0.8438
T,n,w	0.8286	0.6565	0.8656
T,n	0.8093	0.709	0.9081
T,w	0.8026	0.7157	0.9229
n,w	0.7553	0.7893	1.014

In Table 2 "dew point" (c) "relative humidity" (percent), "Sunny Hours" (h), "Saturation vapor pressure deficit" (mbar), "Wind speed" (m/s), "the average monthly temperature '(C) respectively is shown as dwe, RH, n, e, W, T.

Sensitivity analysis

To determine the most important factor for modeling "average monthly potential evapotranspiration for the next month" via Bagging 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 third 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 tree model Bagging. Then the six parameters are located in the first row and are including the parameter "average monthly temperature", "sun hours", "dew point", "relative humidity", "wind speed" and "Saturation vapor pressure deficit" and the five parameters are located in second row are including parameters' average monthly temperature "," wind speed "," relative humidity ": "Saturation vapor pressure deficit" and "Sunny Hours"are the second and third ranks respectively in the positive impact Bagging model to proper functioning in the estimating average monthly evapotranspiration for the next month.

CONCLUSION

From this study it can be concluded that: Techniques of "data mining" such Weka software models can be used to estimate evapotranspiration potential of the next month. Bagging model with an estimate of monthly potential evapotranspiration for the next month is shown that can have a high capacity to estimating meteorological parameters. This model can be used to estimate "potential evapotranspiration" used in a variety of stations that are deficient in recorded meteorological parameters. It was concluded from Table 2: Sensitivity to model weather Bagging enter the six parameters, including «average temperature» (c), «sunny hours" (h), "dew point temperature» (c), «The average relative humidity" (%), " The average speed wind "(meters per second) 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

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

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

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

Jensen ME, Burman RD, Allen RG. 1990. Evapotranspiration and irrigation water requirements. ASCE Manual and Reports on Engineering Practice No. 70. ASCE: New York

Mamashkani A, Nazemi AR. 2009. Introduction to "data mining". Neishabour branch of Islamic Azad University Press , Neishabour.

Quinlan JR. 1992. Learning with continuous classes. Proceeding of Australian Joint Conference on Artificial Intelligence. World Scientific Press: Singapore.

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