World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:1, No:2, 2007

Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values

S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias

***Abstract*—**Estimates of temperature values at a specific time of day, from daytime and daily profiles, are needed for a number of environmental, ecological, agricultural and technical applications, ranging from natural hazards assessments, crop growth forecasting to design of solar energy systems. The scope of this research is to investigate the efficiency of data mining techniques in estimating minimum, maximum and mean temperature values. For this reason, a number of experiments have been conducted with well-known regression algorithms using temperature data from the city of Patras in Greece. The performance of these algorithms has been evaluated using standard statistical indicators, such as Correlation Coefficient, Root Mean Squared Error, etc.

***Keywords*—**regression algorithms, supervised machine learning.

I. Introduction

WEATHER data are generally classified as either synoptic data or climate data. Synoptic data is the real time data provided for use in aviation safety and forecast modelling. Climate data is the official data record, usually provided after some quality control is performed on it. Special networks also exist in many countries that may be used in some cases to provide supplementary climate data.

Knowledge of meteorological data in a site is essential for meteorological, pollution and energy applications studies and development. Especially temperature data is used to determine thermal behaviour (thermal and cooling loads, heat losses and gains) of buildings [2]. It is also an explicit requirement for sizing studies of thermal [13] and/or PV systems [18], [5]. Another major sector where temperature data is fundamental is the estimation of biometeorological parameters in a site [16]. In advanced energy system designs the profile of any meteorological parameter is a prerequisite for systems operating management on daily and/or hourly basis. Also, simulations of long-term performance of energy plants require detailed and accurate meteorological data as input. This knowledge may be obtained, either by the elaboration of data banks, or by the use of estimation methodologies and techniques, where no detailed data are available. As nowadays "smart buildings" have became a reality, artificial techniques must be embedded is building management systems (BMS), in order energy profile (loads, gains etc) of a following time period (next hour, next day) to be predetermined. That will lead to a more effective energy management of the building or the energy plant. Weather data from automated weather stations have also become an important component for prediction and decision making in agriculture and forestry. The data collected from such stations are used in predictions of insect and disease damage in crops, orchards, turfgrasses, and forests [4]; in deciding on crop-management actions such as irrigation [1]; in estimating the probability of occurrence of forest fires [9]; and in many other applications [19].

The scope of this research is to investigate the efficiency of data mining techniques in estimating minimum, maximum and average temperature values. A number of experiments have been conducted with well-known regression algorithms using temperature data from the city of Patras in Greece. The performance of these algorithms has been evaluated using standard statistical indicators.

The following section describes the data set of our study. Section III presents the experimental results for the representative regression algorithms. Finally, section IV discusses the conclusions and some future research directions.

II. Description of our Dataset

The values of temperature data used in this paper were obtained from the meteorological station of the Laboratory of Energy and Environmental Physics of the Department of Physics of University of Patras. Collected data cover a four years period (2002-2005). This station records temperature, relative humidity and rainfall data on hourly basis (8760 measurements per year). For the needs of this work minimum, maximum and average temperature values for the city of Patras were calculated, from the elaboration of the data bank of that station. The minimum, maximum and average temperature values values were inserted in new data banks with reference to the day of the year (D) (1-365). The data were also elaborated per month. In that case average daily temperatures were registered with reference to the number of

S. Kotsiantis is with the Department of Computer Science and Technology, University of Peloponnese, Greece (phone: +302610-997833; fax: +302610-997313; e-mail: sotos@math.upatras.gr).

S. Lykoudis is with the National Observatory of Athens, Institute for Environmental Research and Sustainable Developement, GR-15236 Palia Pendeli, Greece.

A. Argiriou is with the University of Patras, Department of Physics, Section of Applied Physics, GR-26500 Patras, Greece.

the month (1-12), the number of the day of the month (1-30) and finally to the day of the year (D) (1-365).

Many methods have been proposed so far worldwide for the estimation-prediction of monthly, daily or even hourly values of different meteorological parameters [10], [11], [12], [14], based mainly on past time data analysis. Such a simple method is the one proposed by [15]. This method is the result of the elaboration of temperature measurements made by the Hellenic National Meteorological Service (HNMS) in different sites of Greece. The analysis of this data shows that the yearly variation of the average, maximum and minimum values of daily temperature can be expressed by the following equation [15]:

$$T(D) = A + B\sin(\frac{360}{365}D - f) \quad (1)$$

where D is the day of the year (1-365), A is the average yearly temperature in ^{o}C, B is the width of the yearly temperature variation in ^{o}C and f is the phase shift expressed in degrees or days. These variables are typical and have constant value depending on the site of the country. Their values have been calculated for a number of Greek cities using the least square method. As far as Patras is concerned their values for the calculation of average daily temperature are given in the table below (elaboration of temperature data of the period 1960-1974). The parameters of eq(1) have also been re-estimated using the 2002-2005 data.

TABLE 1:VALUES OF A,B AND F FOR THE CITY OF PATRA

	Based on 1960-1974 data	Based on 2002-2005 data
A	17,339	18,351
B	-7,47	-8,65
f	-59,691	-62,908
Correlation coefficient	0.8872	0.8881

III. DATA MINING ALGORITHMS USED

The problem of regression in data mining consists in obtaining a functional model that relates the value of a target continuous variable *y* with the values of variables $x_1, x_2, ..., x_n$ (the predictors). This model is obtained using samples of the unknown regression function. These samples describe different mappings between the predictor and the target variables.

The traditional approach for prediction of a continuous target is the classical linear least-squares regression (LR) [7]. The model constructed for regression in this traditional approach is a linear equation. By estimating the parameters of this equation with a computationally simple process on the training set, a model is created. However, the linearity assumption between input features and predicted value introduces a large bias error for most domains. That is why most studies are directed to nonlinear and, non-parametric techniques for the regression problem.

For the aim of our comparison the most common regression techniques namely Model Trees and Rules [20], instance based learners [3], Artificial Neural Networks, and additive regression [8] are used.

Model trees are the counterparts of decision trees for regression tasks. Model trees are trees that classify instances by sorting them based on attribute values. Instances are classified starting at the root node and sorting them based on their attribute values. The most well known model tree inducer is the M5´ [20]. A model tree is generated in two stages. The first builds an ordinary decision tree, using as splitting criterion the maximization of the intra-subset variation of the target value [20]. The second prunes this tree back by replacing subtrees with linear regression functions wherever this seems appropriate.

M5rules algorithm produces propositional regression rules in IF-THEN rule format using routines for generating a decision list from M5´Model trees [21]. The algorithm is able to deal with both continuous and nominal variables, and obtains a piecewise linear model of the data.

Instance-based learning algorithms are lazy-learning algorithms, as they delay the induction or generalization process until regression process is performed. k-Nearest Neighbour (kNN) is based on the principle that the instances within a dataset will generally exist in close proximity with other instances that have similar properties [3]. KNN algorithm first finds the closest instances to the query point in the instance space according to a distance measure, and then outputs the average of the target values of those instances as the prediction for that query instance. As the prediction of the target value of a query instance requires to measure its distance to all training instances, which might be a very huge set, the prediction in KNN is very costly. IB3 is a well known technique for instance based learning.

Artificial Neural Networks (ANNs) are another method of inductive learning based on computational models of biological neurons and networks of neurons as found in the central nervous system of humans [17]. Regression with a neural network takes place in two distinct phases. First, the network is trained on a set of paired data to determine the input-output mapping. The weights of the connections between neurons are then fixed and the network is used to predict the numerical class values of a new set of data. The most well-known learning algorithm to estimate the values of the weights of a neural network - the Back Propagation (BP) algorithm [17] - was the representative of the Neural Networks.

Combining models is not a really new concept for the statistical pattern recognition, machine learning, or engineering communities, though in recent years there has been an explosion of research exploring creative new ways to combine models. Currently, there are two main approaches to model combination. The first is to create a set of learned models by applying an algorithm repeatedly to different training sample data; the second applies various learning algorithms to the same sample data. The predictions of the models are then combined according to an averaging scheme.

A method that uses different subset of training data with a single learning method is the boosting approach [6]. The boosting approach uses the base models in sequential collaboration, where each new model concentrates more on

the examples where the previous models had high error. Although boosting for regression has not received nearly as much attention as boosting for classification, there is some work examining gradient descent boosting algorithms in the regression context. Additive regression [8] is a well known boosting method for regression.

IV. Results

For the regression methods, there isn't only one regressor's criterion. Table 2 represents the most well known. Fortunately, it turns out for in most practical situations the best regression method is still the best no matter which error measure is used.

In order to calculate the models' regressor criteria for our experiments, we used the free available source code for most of the algorithms by [21] for our experiments.

TABLE II: REGRESSOR CRITERIA (P: PREDICTED VALUES, A: ACTUAL VALUES)

Correlation coefficient	$R=\frac{S_{PA}}{\sqrt{S_P S_A}}$ where $S_{PA}=\frac{\sum_i (p_i-\bar{p})(a_i-\bar{a})}{n-1}$, $S_P=\frac{\sum_i (p_i-\bar{p})^2}{n-1}$, $S_A=\frac{\sum_i (a_i-\bar{a})^2}{n-1}$
Root mean squared error	$\sqrt{\frac{(p_1-a_1)^2+\ldots\ (p_1\ \ a_1)^2}{n}}$

In the following three tables we present the models' regressor criteria in predicting average daily temperature values using as input a) the previous year data (2004) in Table 3; b) the last two years (2003,2004) in Table 4 and c) the three last years (2002,2003,2004) in Table 5.

TABLE III: PREDICTING AVERAGE DAILY TEMPERATURE VALUES USING AS INPUT THE PREVIOUS YEAR DATA (2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9288	2.6131
M5rules	0.9233	2.7392
Additive Regression (50 iterations)	0.9279	2.5698
LR	0.8529	5.6715
IB3	0.9002	3.0715
BP	0.92	3.2724

TABLE IV: PREDICTING AVERAGE DAILY TEMPERATURE VALUES USING AS INPUT THE LAST TWO YEARS DATA (2003-2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9442	2.357
M5rules	0.9417	2.4309
Additive Regression (50 iterations)	0.9319	2.6129
LR	0.8529	5.2118
IB3	0.9195	2.8602
BP	0.9099	3.7817

TABLE V: PREDICTING AVERAGE DAILY TEMPERATURE VALUES USING AS INPUT THE LAST THREE YEARS DATA (2002-2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9336	2.5369
M5rules	0.9365	2.4557
Additive Regression (50 iterations)	0.9288	2.6334
LR	0.8529	5.5012
IB3	0.9205	2.7801
BP	0.9097	4.4836

As a result, the experts are in the position using the temperatures of previous years, to predict average daily temperature values of the examined year with sufficient precision, which reaches 92% correlation coefficient in the initial forecasts (using the data of the previous of the examined year) and exceeds the 94% using the data of the last two years before the examined year.

In the following three tables we present the models' regressor criteria in predicting minimum daily temperature values using as input a) the previous year data (2004) in Table 6; b) the last two years (2003,2004) in Table 7 and c) the three last years (2002,2003,2004) in Table 8.

TABLE VI: PREDICTING MINIMUM DAILY TEMPERATURE VALUES USING AS INPUT THE PREVIOUS YEAR DATA (2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9437	2.2219
M5rules	0.9147	2.5995
Additive Regression (50 iterations)	0.9315	2.3552
LR	0.8648	5.2159
IB3	0.897	2.88
BP	0.9216	2.8999

TABLE VII: PREDICTING MINIMUM DAILY TEMPERATURE VALUES USING AS INPUT THE LAST TWO YEARS DATA (2003-2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.945	2.1386
M5rules	0.9424	2.182
Additive Regression (50 iterations)	0.9317	2.3735
LR	0.8648	4.814
IB3	0.9217	2.5578
BP	0.9137	3.9315

TABLE VIII: PREDICTING MINIMUM DAILY TEMPERATURE VALUES USING AS INPUT THE LAST THREE YEARS DATA (2002-2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9395	2.246
M5rules	0.9391	2.2417
Additive Regression (50 iterations)	0.9322	2.3818
LR	0.8648	5.0424
IB3	0.9207	2.5569
BP	0.9137	4.1722

As a result, the experts are in the position using the temperatures of previous years, to predict minimum daily temperature values of the examined year with sufficient precision, which reaches 93% correlation coefficient in the initial forecasts (using the data of the previous of the examined year) and exceeds the 94% using the data of the last two years before the examined year.

In the following three tables we present the models' regressor criteria in predicting maximum daily temperature values using as input a) the previous year data (2004) in Table 9; b) the last two years (2003,2004) in Table 10 and c) the three last years (2002,2003,2004) in Table 11.

TABLE IX: PREDICTING MAXIMUM DAILY TEMPERATURE VALUES USING AS INPUT THE PREVIOUS YEAR DATA (2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9053	3.2671
M5rules	0.8959	3.4213
Additive Regression (50 iterations)	0.9044	3.1572
LR	0.8231	6.4537
IB3	0.873	3.7743
BP	0.8958	4.0265

TABLE X: PREDICTING MAXIMUM DAILY TEMPERATURE VALUES USING AS INPUT THE LAST TWO YEARS DATA (2003-2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9196	3.0779
M5rules	0.9168	3.0888
Additive Regression (50 iterations)	0.9119	3.2432
LR	0.8231	5.9401
IB3	0.8892	3.6713
BP	0.8826	7.5053

TABLE XI: PREDICTING MAXIMUM DAILY TEMPERATURE VALUES USING AS INPUT THE LAST THREE YEARS DATA (2002-2004)

	Correlation coefficient	Root mean Squared error (°C)
M5	0.9165	3.1119
M5rules	0.9163	3.1261
Additive Regression (50 iterations)	0.9077	3.2618
LR	0.8231	6.2729
IB3	0.8903	3.5475
BP	0.7942	7.7678

As a result, the experts are in the position using the temperatures of previous years, to predict maximum daily temperature values of the examined year with sufficient precision, which reaches 90% correlation coefficient in the initial forecasts (using the data of the previous of the examined year) and exceeds the 92% using the data of the last two years before the examined year.

As a general conclusion, it was found that the regression algorithms could enable experts to predict minimum, maximum and average temperature values with satisfying accuracy using as input the temperatures of the previous years. We believe that using as input the temperatures of the two

previous years gives sufficient results. There is no need for more historical data.

V. Conclusion

Ideally, the market needs timely and accurate weather data. In order to achieve this, data should be continuously recorded from stations that are properly identified, manned by trained staff or automated with regular maintenance, in good working order and secure from tampering. The stations should also have a long history and not be prone to relocation. The collection and archiving of weather data is important because it provides an economic benefit but the local/national economic needs are not as dependent on high data quality as is the weather risk market.

In this study, it was found that the regression algorithms could enable experts to predict minimum, maximum and average temperature values with satisfying accuracy using as input the temperatures of the previous years. The methods used in this work, for the case of Patras, should be tested and in other regions with different climatic profile. Also, other methodologies (fuzzy logic techniques etc) have to be validated in many regions of the country covering its climatic spectrum, including not only temperature data (on any time basis) but other meteorological parameters as well (wind speed, solar radiation etc).

References

[1] Acock M. C., Pachepsky Ya. A., Estimating Missing Weather Data for Agricultural Simulations Using Group Method of Data Handling, Journal of Applied Meteorology: Vol. 39, No. 7, pp. 1176–1184, 2000.
[2] Ashrae, Handbook of Fundamentals, American Society of Heating, Refrigerating and Air Conditioning Engineers, New York: 1993
[3] Atkeson, C. G., Moore, A.W., & Schaal, S., Locally weighted learning. Artificial Intelligence Review, 11, (1997) 11–73.
[4] Dinelli, D., 1995: What weather stations can do. Landscape Manage., 34 (3), 6G.
[5] Duffie, J.A., and W.A Beckman. 1991. Solar Engineering of thermal processes. New York: John Wiley and Sons
[6] Duffy, N. Helmbold, D., Boosting Methods for Regression, Machine Learning, 47, (2002) 153–200.
[7] Fox, J. (1997), Applied Regression Analysis, Linear Models, and Related Methods, ISBN: 080394540X, Sage Pubns.
[8] Friedman J. (2002). "Stochastic Gradient Boosting," Computational Statistics and Data Analysis 38(4):367-378.
[9] Fujioka, F. M., 1995: High resolution fire weather models. Fire Manage. Notes, 57, 22–25.
[10] Gelegenis, J.J. 1999. 'Estimation of hourly temperature data from their month average values: case study of Greece.' Renewable Energy 18, nos 1: 49-60
[11] Hall, I.J., Generation of a Typical Meteorological Year, Proceedings of the 1978 annual meeting af AS of ISES, Denver USA, 1979
[12] Jain, P.C., Comparison of techniques for the estimation of daily global irradiation and a new model for the estimation of hourly global irradiation. Solar and Wind Technology 1, nos. 2, 1984, pp.123-134
[13] Klein, S.A, W.A Beckman and J.A. Duffie. 1985. 'A Design Procedure for Solar Heating systems.' Solar Energy 18: 113-127.
[14] Knight, K.M., Klein, S.A and Duffie, J.A., A methodology for the synthesis of hourly weather data. Solar Energy 46, nos 2, 1991, pp.109-120.
[15] Kouremenos D.A, Antonopoulos K.A, Temperature data for 35 Greek cities. In Greek. Athens 1993 – Second Edition.
[16] Matzarakis, A. 1995. Human-biometeorological assessment of the climate of Greece. Ph.D. Dissertation, University of Thessaloniki.
[17] Mitchell, T., Machine Learning, McGraw Hill, 1997.
[18] Rahman S. and Chowdhury B., "Simulation of Photovoltaic power systems and their performance prediction". IEEE Transactions on Energy Conversion 3,440-446 (1988)
[19] Tugay Bilgin and Yilmaz Çamurcu, 2004, A Data Mining Application on Air Temperature Database, in LNCS 3261 - Advances in Information Systems, Springer Berlin / Heidelberg, ISBN 978-3-540-23478-4, pp.68-76
[20] Wang, Y. & Witten, I. H., Induction of model trees for predicting continuous classes, In Proc. of the Poster Papers of the European Conference on ML, Prague (pp. 128–137).
[21] Witten, I.H., Frank, E., "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Sotiris Kotsiantis received a diploma in mathematics, a Master and a Ph.D. degree in computer science from the University of Patras, Greece. He is an adjunct lecturer in the Department of Computer Science and Technology at the University of Peloponnese, Greece His main research interests are in the field of machine learning, data mining and knowledge representation. He has about 80 publications to his credit in international journals and conferences.