A Hybrid Data Mining Technique for Estimating Mean Daily Temperature Values

S. Kotsiantis[1], A. Kostoulas[2], S. Lykoudis[3,4], A. Argiriou[3], K. Menagias[5]

[1] *Educational Software Development Laboratory, University of Patras, Greece*

[2] *Mechanical Engineer T.E, M.Sc*

[3] *University of Patras, Department of Physics, Section of Applied Physics, GR-26500 Patras, Greece*

[4] *National Observatory of Athens, Institute for Environmental Research and Sustainable Developement, GR-15236 Palia Pendeli, Greece*

[5] *Mechanical Engineer T.E*

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 initial scope of this research is to investigate the efficiency of data mining techniques in estimating mean daily 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. Finally, a hybrid data mining technique is proposed that can be used to predict more accurately the mean daily temperature values.

Keywords: regression algorithms, supervised machine learning.

1. 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.

The scope of this research is to investigate the efficiency of data mining techniques in estimating mean daily 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. Finally, a hybrid data mining technique is proposed that can be used to predict more accurately the mean daily temperature values.

The following section describes the applications of meteorological data and the data set of our study. Section 3 presents the experimental results for the representative regression algorithms and the proposed hybrid data mining technique. Finally, section 4 discusses the conclusions and some future research directions.

2. *Applications of the Meteorological Data and Description of our Dataset*

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 [Ashrae (1993)]. It is also an explicit requirement for sizing studies of thermal [Klein et. Al. (1995)] and/or PV systems [Rahman and Chowdhury (1988)], [Duffie and Beckman (1991)]. Another major sector where temperature data is fundamental is the estimation of biometeorological parameters in a site [Matzarakis (1995)]. 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 [Dinelli (1995)]; in deciding on crop-management actions such as irrigation [Acock and Pachepsky (2000)]; in estimating the probability of occurrence of forest fires [Fujioka (1995)]; and in many other applications.

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 mean daily temperature values for the city of Patras were calculated, from the elaboration of the data bank of that station. The mean daily temperature values were inserted in a new data bank with reference to the day of the year (D) (1-365). The data were also elaborated per month. In that case mean daily temperatures were registered with reference to the number of 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 [Gelegenis (1999)], [Hall (1979)], [Jain (1984)], [Knight et. Al. (1991)], based mainly on past time data analysis. Such a simple method is the one proposed by [Kouremenos and Antonopulos (1993)]. 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 mean, maximum and minimum values of daily temperature can be expressed by the following equation [Kouremenos and Antonopulos (1993)]:

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

where D is the day of the year (1-365), A is the mean 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 mean 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

3. 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.

For the aim of our comparison the most common regression techniques namely Model Trees and Rules [Wang and Witten (1997)], instance based learners [Atkeson et. Al. (1997)] and additive regression [Friedman (2002)] 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´ [Wang and Witten (1997)]. 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 [Wang and Witten (1997)]. 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 [Witten and Frank (2000)]. The algorithm is able to deal with both continuous and nominal variables, and obtains a piecewise linear model of the data.

Locally weighted linear regression (LWR) is a combination of instance-based learning and linear regression [Atkeson et. Al. (1997)]. Instead of performing a linear regression on the full, unweighted dataset, it performs a weighted linear regression, weighting the training instances according to their distance to the test instance at hand. This means that a linear regression has to be performed for each new test instance, which makes the method computationally quite expensive. However, it also makes it highly flexible, and enables it to approximate non-linear target functions. K* is a well known technique for instance based learning.

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 [Duffy and Helmbold (2002)]. 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 [Friedman (2002)] is a well known boosting method for regression.

It is well-known that the selection of an optimal set of regression models is an important part of multiple regression model systems and the independence of regression model outputs is generally considered to be an advantage for obtaining better multiple regression model systems. The proposed heterogeneous ensemble method relies on the idea of selection in ensemble creation and combines the advantages of regression model fusion and dynamic selection, which are the two main categories of traditional combining algorithms. In terms of regression model combination, the averaging methods demand no prerequisites from the regression models.

4.1 Proposed Algorithm

The motivation of our approach follows from a key observation regarding the bias/variance decomposition, namely the fact that ensemble averaging does not affect the bias portion of the error, but reduces the variance, when the estimators on which averaging is done are independent. It is easy to come up with examples where simple averaging is optimal. For example if all estimators are unbiased and uncorrelated with identical variances or in general, when symmetry indicates that no single estimator should be preferred.

The presented methodology is six steps strategy:

- The dataset is sampled at random about 20% of the initial set
- The new dataset is divided at random into three equal parts
- Two of three parts are used for training of algorithms and the remaining data is the testing set
- The result of three tests are averaged
- The algorithms that have statistically worse performance (according to t-test with $p<0.05$) than the most accurate are not used by the ensemble
- The remaining algorithms then executes on the full training set to produce the prediction model using the averaging rule.

In detail, the regression process includes two phases: (1) *learning phase*, and (2) *application phase*. During the learning phase, a set of base regression models is generated and each base regression model in the ensemble (regression models h_1 … h_n) is trained. In order to improve the generalization capabilities of the aggregate predictor one must generate diverse individual predictors retaining only those that

perform reasonably well on dataset. This can be accomplished by the following procedure during the learning phase, which is described in Table 2 as a general pseudo-code.

Table 2. *The Learning Process of the Proposed Ensemble (Selective Averaging)*

//Learning Phase

1. Input Dataset D
2. Select Regression Models h_1 h_2 … h_n
3. D´= random sample of D // 20% of D
4. Randomly split D´ into three equal parts (D_1, D_2, D_3)
5. For k=1 to n do
   { For i=1 to 3 do
      {Compute Training set $Tr_i = (D- D_i)$
       Train algorithm h_k using Tr_i
       Evaluate the performance P_i of h_k using D_i as test examples
      } //end for i
   $Ph_k = (P_1 + P_2 + P_3)/3$ // true performance of h_k
   }//end for k
   // Find best regression model h_{Best}
6. Max= Ph_1
7. Best = 1
8. For l = 2 to N
   {If (Ph_l >Max) Then
      {Max = Ph_l
      Best = l}//endif
   }//end for l
9. t=0
10. For m = 1 to N
    {If (h_m is not statistically worse than h_{Best} according to paired t-test with p-value<0.05) then
       {t=t+1
       Train algorithm $h'_t=h_m$ using D as training Set}//endif

During the application phase, which is described in Table 3 as a general pseudo-code, the corresponding predictions of the selected base regression models are combined with averaging rule to produce the final decision of the ensemble. Thus, the final prediction is given as $y^*=\sum_{i=1}^{t}\frac{h_i}{t}$, where t is the number of the selected base regression models by the proposed technique.

Table 3. *The Application Phase of the Proposed Ensemble (Selective Averaging)*

//Application Phase
1. Input Test Instance A
2. Sum=0
3. For i=1 to t {Sum= Sum+ h'_i (A)} //end for i
4. y* = Sum/t //Final prediction
5. display y*

It must also be mentioned that the proposed ensemble can easily be parallelized using a learning algorithm per machine. Parallel and distributed computing is of most importance for machine learning practitioners because taking advantage of a parallel or a distributed execution a machine learning system may: i) increase its speed; ii) increase the range of applications where it can be used (because it can process more data, for example).

4.2 Results

For the regression methods, there isn't only one regressor's criterion. Table 4 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 [Witten and Frank (2000)] for our experiments. In the following tables we present the models' regressor criteria using as input a) the previous year data (2004) in Table 5; b) the last two years (2003,2004) in Table 6 and c) the three last years (2002,2003,2004) in Table 7.

Table 4. *Regressor criteria (p: predicted values, a: actual values)*

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

Table 5. *Using as input the previous year data (2004)*

	M5	M5rules	Additive Regression (50 iterations)	K*	Simple Procedure (described in Section 2)	Selective Averaging
Correlation coefficient	0.9189	0.914	0.9153	0.9256	0.8882	0.9316
Root mean Squared error (°C)	2.6454	2.7308	2.6465	2.6977	2.2044	2.627

Table 6. *Using as input the last two years data (2003-2004)*

	M5	M5rules	Additive Regression (50 iterations)	K*	Simple Procedure (described in Section 2)	Selective Averaging
Correlation coefficient	0.9331	0.931	0.9156	0.9202	0.8880	0.9413
Root mean Squared error (°C)	2.4635	2.4994	2.7352	2.8006	2.267	2.417

Table 7. *Using as input the last three years data (2002-2004)*

	M5	M5rules	Additive Regression (50 iterations)	K*	Simple Procedure (described in Section 2)	Selective Averaging
Correlation coefficient	0.927	0.9275	0.9178	0.9275	0.8881	0.9501
Root mean Squared error (°C)	2.483	2.4376	2.6266	2.8234	2.2233	2.327

As a result, the experts are in the position using the temperatures of previous years, to predict temperature values of the examined year with sufficient precision, which reaches 91% correlation coefficient in the initial forecasts (using the data of the previous of the examined year) and exceeds the 95% using the data of the last three years before the examined year. It must be mentioned that the proposed hybrid data mining technique produce the most accurate results.

4. 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 temperature values with satisfying accuracy using as input the temperatures of the previous years. It must be mentioned that the proposed hybrid data mining technique produce the most accurate results. The next phase of this work is the implementation and validation of the techniques analyzed and validated here, using minimum and maximum daily temperatures data. 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 (like Neural Networks, 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).

5. References

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.

Ashrae, Handbook of Fundamentals, American Society of Heating, Refrigerating and Air Conditioning Engineers, New York: 1993

Atkeson, C. G., Moore, A.W., & Schaal, S., Locally weighted learning. Artificial Intelligence Review, 11, (1997) 11–73.

Dinelli, D., 1995: What weather stations can do. Landscape Manage., 34 (3), 6G.

Duffie, J.A., and W.A Beckman. 1991. Solar Engineering of thermal processes. New York: John Wiley and Sons

Duffy, N. Helmbold, D., Boosting Methods for Regression, Machine Learning, 47, (2002) 153–200.

Friedman J. (2002). "Stochastic Gradient Boosting," Computational Statistics and Data Analysis 38(4):367-378.

Fujioka, F. M., 1995: High resolution fire weather models. *Fire Manage. Notes,* **57,** 22–25.

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

Hall, I.J., Generation of a Typical Meteorological Year, Proceedings of the 1978 annual meeting af AS of ISES, Denver USA, 1979

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

Klein, S.A, W.A Beckman and J.A. Duffie. 1985. 'A Design Procedure for Solar Heating systems.' Solar Energy 18: 113-127.

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.

Kouremenos D.A, Antonopoulos K.A, Temperature data for 35 Greek cities. In Greek. Athens 1993 – Second Edition.

Matzarakis, A. 1995. Human-biometeorological assessment of the climate of Greece. Ph.D. Dissertation, University of Thessaloniki.

Rahman S. and Chowdhury B., "Simulation of Photovoltaic power systems and their performance prediction". IEEE Transactions on Energy Conversion 3,440-446 (1988)

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).

Witten, I.H., Frank, E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Mateo, CA, (2000).