Influence of kNN-Based Load Forecasting Errors on Optimal Energy Production

Alicia Troncoso Lora[1], José C. Riquelme[1], José Luís Martínez Ramos[2],
Jesús M. Riquelme Santos[2], and Antonio Gómez Expósito[2]

[1] Department of Languages and Systems, University of Sevilla, Sevilla, Spain
{ali,riquelme}@lsi.us.es
[2] Department of Electrical Engineering, University of Sevilla, Sevilla, Spain
{camel,jsantos,age}@us.es

Abstract. This paper presents a study of the influence of the accuracy of hourly load forecasting on the energy planning and operation of electric generation utilities. First, a *k* Nearest Neighbours (kNN) classification technique is proposed for hourly load forecasting. Then, obtained prediction errors are compared with those obtained results by using a M5'. Second, the obtained kNN-based load forecast is used to compute the optimal on/off status and generation scheduling of the units. Finally, the influence of forecasting errors on both the status and generation level of the units over the scheduling period is studied.

Keywords. Nearest neighbours, load forecasting, optimal energy production.

1 Introduction

The prediction of future loads is crucial for the economic and secure operation of electrical power systems. In the short, medium and long term, generation scheduling comprises a set of interrelated optimization problems that require a load forecasting procedure. Consequently, accurate forecasting techniques are crucial for the electric power industry to reduce the uncertainty of the load and to compute an optimal and realistic generation scheduling.

Nowadays, forecasting methods for load estimation can be classified in two main groups: classical statistical methods and techniques based on machine learning. Classical statistical methods [1,2] aim at estimating the current load from the values of past load. The relationships between the load and other relevant factors (e.g., temperature) are used to determine the underlying model of the load time series, the main advantage of classical methods being their inherent simplicity. However, as the relationships between load and factors which have influence on load are nonlinear, it is not an easy task to identify realistic and accurate models using classical methods.

In the last years, techniques based on machine learning such as Artificial Neural Networks (ANN) [3,4] have been applied to one day-ahead load forecasting. The ANNs are trained to learn the relationships between the input variables

Fernando Moura Pires, Salvador Abreu (Eds.): EPIA 2003, LNAI 2902, pp. 189–203, 2003.

(mainly preceding loads and actual temperature) and historical load patterns. The main disadvantage of ANNs is the required learning procedure.

More recently, classification techniques based on the nearest neighbours have been successfully applied in different areas from the traditional pattern recognition such as medical diagnosis tools, game theory expert systems or time series forecasting. Several papers have been published on the application of nearest neighbours techniques to the electricity market price forecasting [6,7], but applications to load forecasting problems are missed.

This paper presents a study of the effects of the accuracy of hourly load forecasting on the generation planning and operation of an electric utility. First, a kNN classification technique is proposed for load forecasting. Then, the hourly load forecasting errors are compared with those obtained results by using a M5'. Second, the obtained kNN-based load forecasts are used to compute the optimal on/off status and generation scheduling of the units. Then, the influence of forecasting errors on both the status and generation level of the units over the scheduling period is studied.

2 One Day-Ahead Load Forecasting

The one day-ahead load forecasting problem aims at predicting the load for the twenty-four hours of the next day. To solve this problem two schemes can be considered:

1) Iterative Scheme: This scheme aims at predicting the load of one hour and the obtained prediction is used as an input for the load forecasting of the next hour. The process is repeated until the load forecasting of the next 24 hours is obtained. The iterative prediction has the disadvantage that the errors are accumulated throughout the prediction horizon.
2) Direct Scheme: This scheme aims at predicting the next twenty-four hours from the same input data. The direct prediction does not take into account the relationships between the load of one hour and the load of successive hours.

Test results have shown similar accuracy for both schemes, and, consequently, the direct scheme has been adopted in this study.

2.1 Description of the Proposed Approach

In this section, an algorithm based on kNN [8] for hourly load forecasting is described. kNN algorithms are techniques for pattern classification based on the similarity of the individuals of a population. The members of a population coexist surrounded of similar individuals which have similar properties. This simple idea is the learning rule of the kNN classifier. Thus, the nearest neighbours' decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points. In contrast to statistical methods that try to identify a model from the available data, the kNN method uses the training set as the model.

A particular kNN algorithm is characterized by issues such as the number of neighbours, type of distance used, etc.

In the method used in this paper, each individual is defined by the 24-hours load of a day. Thus, the kNN classifier finds the daily load curve that is "similar to" the load curve of previous days.

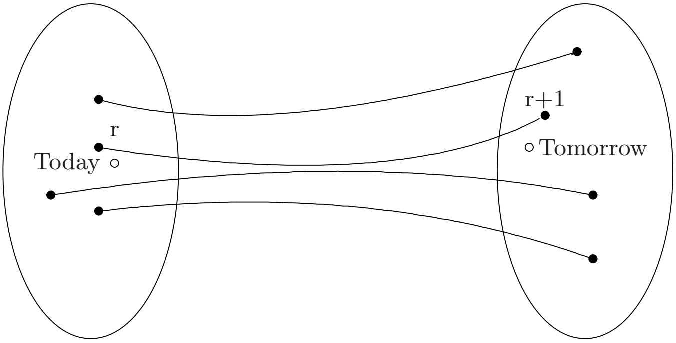

Fig. 1. Learning rule of the proposed approach.

The basic algorithm for the prediction of the electric energy demand for day $d+1$ can be written as follows:

1. Calculate the distances between the load of the day d, D_d, and the preceding points $\{D_{d-1}, D_{d-2}, ...\}$ using a metric *dist*. Let v_1,...,v_k be the k nearest days to the day d, sorted by closeness.
2. The prediction is:

$$\widehat{D}_{d+1} = \frac{1}{\alpha_1 + ... + \alpha_k} \sum_{j=1}^{k} \alpha_j \cdot D_{v_j+1} \tag{1}$$

where

$$\alpha_j = \frac{dist(D_d, D_{v_k}) - dist(D_d, D_{v_j})}{dist(D_d, D_{v_k}) - dist(D_d, D_{v_1})} \tag{2}$$

Notice that $0 \leq \alpha_j \leq 1$, i.e., the weight is equal to zero when the considered day is the most distant and one when the considered day is the nearest.

Once the forecasted load of the day $d+1$ has been obtained, the actual load of the day $d+1$ is used as an input for the load forecasting of the day $d+2$.

Notice that the prediction aims at estimating the load for a certain day from a linear combination of the load of the days that follow the nearest neighbours days.

If the k nearest neighbours for a vector D_d are $[D_{v_1}, ..., D_{v_k}]$, where v_i is the i^{th} nearest neighbour, the set of points $[D_{v_1+1}, ..., D_{v_k+1}]$ will usually be the nearest to D_{d+1} for noise-free time series.

Figure 1 shows the geometric idea of the kNN classifier when the considered number of neighbours is equal to one. Today's hourly load and the unknown load of tomorrow are represented by circumferences. The four black points are the neighbours of today's load. The point r is the nearest neighbour. Then, a possible estimation for tomorrow's load is the load of the day $r + 1$.

In the classical kNN, the nearest neighbours of tomorrow's load are used for the prediction for tomorrow's load, but this is not possible. Thus, the method presented in this section is the adapted kNN algorithm.

Some key issues of the proposed technique are the following:

- **Choice of a metric:** A time series Y can be considered as a point in a n-dimensional space. Given a sequence query, q, a sequence of Y with the same length as q is searched out, z, such that the distance between the sequences is minimum. The choice of the metric to measure the similarity between two time series depends mainly on the specific features of the considered series. The most common metric is the square of the Euclidean distance, although other metrics can be used [9,10].
- **Number of neighbours:** The accuracy of hourly load forecasting can be influenced by this parameter. In practice, the optimal value of k is usually small for noise-free time series, since only a small number of different values for k must to be considered to find the optimal value. In this paper, k is determined by minimizing the mean relative, absolute and square errors for the training set.

2.2 Numerical Results

The kNN algorithm described in the previous section has been applied in several experiments to obtain the forecast of the Spanish electric energy demand. The working days of the period January 2000-May 2001 have been used to determine the optimal number of neighbours and the distance to measure the similarity between two curves.

The available period of June-November 2001 (Summer-Autumn seasons) has been chosen as a test set to check the forecasting errors and to validate the proposed method.

Figure 2a and 2b show the influence of the number of neighbours used for the next-day load forecasting on the mean relative, absolute and square errors for the considered training set, the distance to evaluate the similarity between a previous day and the historical data being the Euclidean and Manhatan distance, respectively.

From Figure 2, the following conclusions can be stated for the load time series:

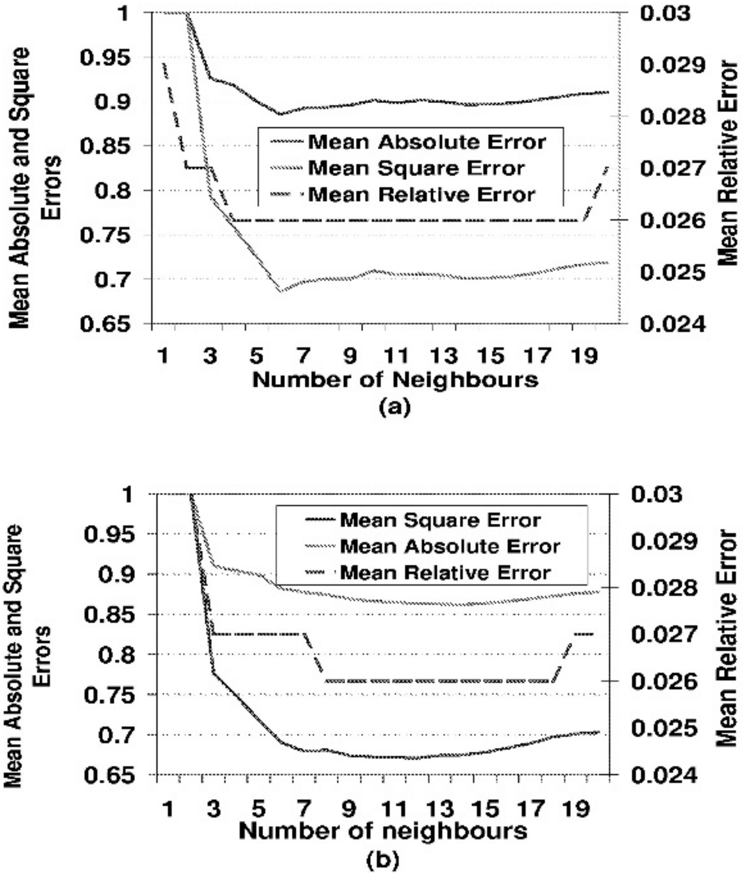

Fig. 2. Optimal number of neighbours using **a)** the Euclidean Distance, **b)** the Manhatan Distance.

1. The optimal number of neighbours is equal to six using the Euclidean distance while it is equal to thirteen using the Manhatan distance. Consequently, this number depends on the type of norm used to compute the distance.
2. The optimal number of neighbours is independent of the type of error used like objective function to minimize. For example, the optimal number of neighbours is six for all type of errors (relative, absolute and square error) when the Euclidean distance is considered.

Test results have shown the same average error for the training set when two distances have been considered: the Euclidean and Manhatan distance. Thus, the Manhatan distance is only considered in the sequel.

Figure 3a shows the hourly average of the real and forecasted load for the working days from June 2001 to November 2001, being the mean error 2.3%. A

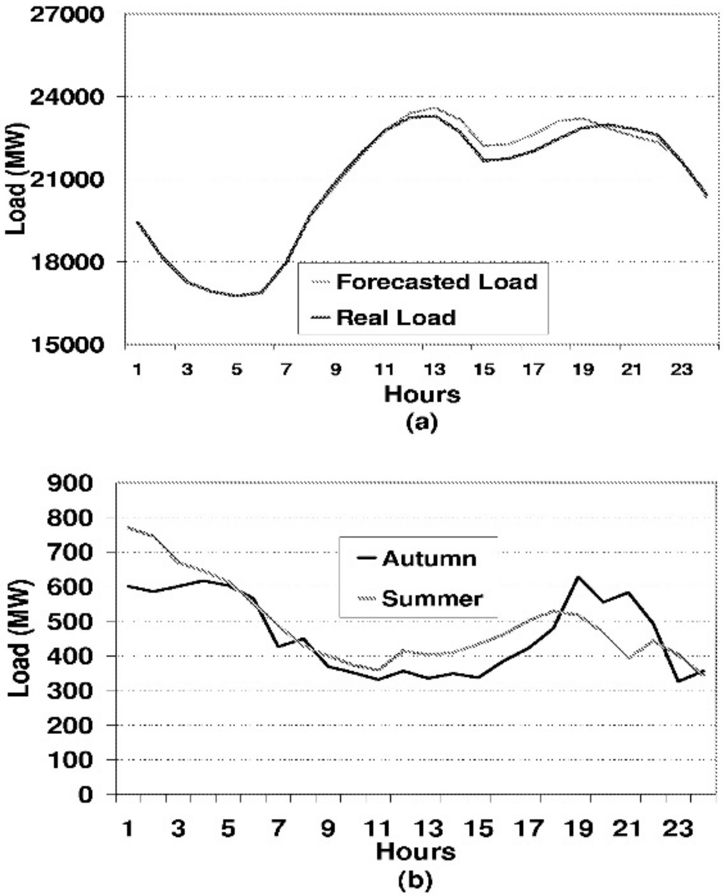

Fig. 3. a) Hourly average of real and forecasted load; **b)** Absolute-value hourly average of the forecasted load error.

good performance of the prediction method based on kNN can be observed. Note that the results obtained applying this method are similar than those carried out using other techniques such as ANNs [5].

Figure 3b presents the hourly average absolute value of the error of the forecasted load for the Autumn and Summer seasons. Note that the forecasting errors are larger during valley hours. However, it is more important to obtain an accurate prediction during peak hours because the electric energy is more expensive during these hours.

Figure 4a and 4b present the forecasted load for the two weeks that lead to the largest and smallest average errors, along with the actual load for the test set. The weeks with the largest and smallest errors correspond to Tuesday September 11th until Monday September 17th, and Monday October 22nd until

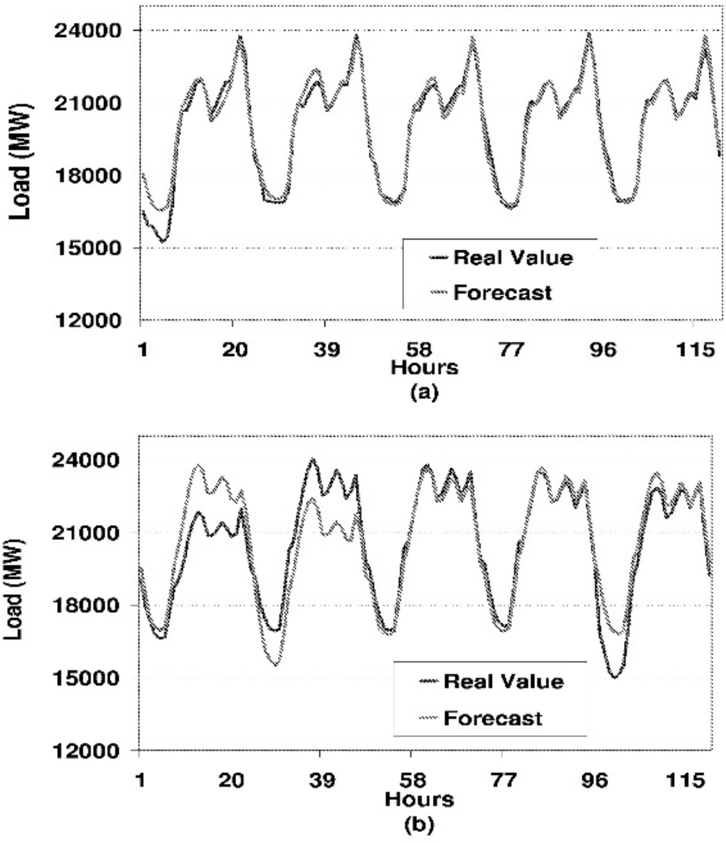

Fig. 4. a) Best weekly forecasts; **b)** Worst weekly forecasts.

Friday October 26th, respectively. It can be observed that the week with the higher prediction errors is the one that corresponds to the terrorist assault of the New York Twin Towers. Notice that the prediction errors corresponding to September 12th are rather high due to the anomalous load of the day before.

Table 1. Daily mean errors of the best and worst weekly forecasts.

Days	1	2	3	4	5	Mean (%)
October 22th-October 26th	3.13	1.25	1.11	0.08	0.07	1.4
September 11th-September 17th	5.88	6.73	1.18	1.13	4.55	3.9

The five working-day mean errors for the best and worst weeks are shown in Table 1. The weekly mean errors are 1.4% and 4%, respectively.

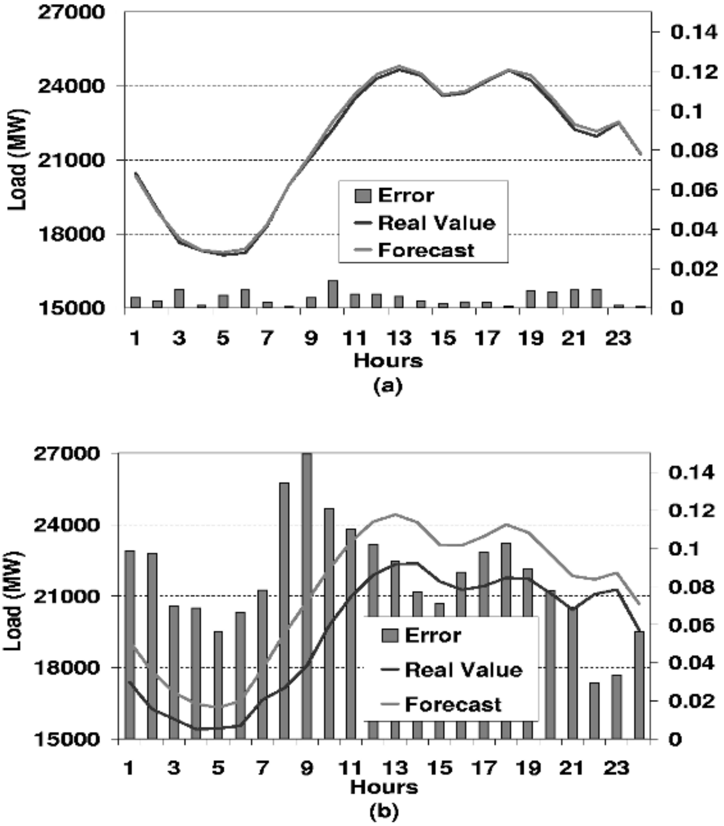

Fig. 5. a) Best daily forecasts; **b)** Worst daily forecasts.

Figure 5a and 5b present the forecasted load for the two days that lead to the largest and smallest average relative errors, along with the actual load for the test set. The days with the largest and smallest errors correspond to Monday August 6th and Tuesday July 17th, respectively. It can be observed that the day with the higher prediction errors is the one that corresponds to the first Monday of August, the day of the beginning of the Summer holiday for most of Spanish people.

The results obtained applying this method are compared with those carried out using the classifier M5' described in the Appendix.

Table 2 presents the mean relative value of forecasting errors and the maximum and minimum daily errors for the test set. Fifty-three rules of the dependent variables are found by M5' for the load forecasting problem. Note that the obtained results applying the kNN are better than those carried out using the M5' algorithm. The average error is 11% when the M5' is used, while the obtained average error using a kNN is 2.3%.

Table 2. Comparison of predicted daily demand for both methods.

	June-November 2001	
	kNN	M5'
Minimum daily errors (%)	0.5	7
Maximum daily errors (%)	8.5	16.5
Average Relative errors (%)	2.3	11

3 The Optimal Energy Production Problem

Once the forecasted demand profile has been obtained, the optimal status on/off and hourly generation level of the units must be determined in order to minimize the expected total cost satisfying the system load forecasting.

This module computes the optimal solution of the classical short-term Unit Commitment and Economic Dispatch (UC-ED) problem [11]. The goal is to obtain, for every hour, the optimal on/off status and generated power of each generating unit so that the total demand is satisfied in the presence of technical constraints.

3.1 Objective Function

The total generation cost of the scheduling period, given the on/off status of the thermal units, $U_{i,t}$, is defined by

$$C_T = \sum_{t=1}^{n_t} \sum_{i=1}^{n_g} \{C_{i,t} \cdot U_{i,t} \ + SU_i \cdot U_{i,t} \cdot (1 - U_{i,t-1}) + SD_i \cdot (1 - U_{i,t}) \cdot U_{i,t-1}\} \quad (3)$$

where n_t is the number of hours of the scheduling period, n_g is the number of thermal units, each having a cost function $C_{i,t} = C_i(P_{i,t})$ of the generated power $P_{i,t}$, and SU_i, SD_i are respectively the start-up and shut-down cost of generator i.

3.2 Constraints

The minimization of the objective function is subject to the following constraints:

- Upper and lower generation limits of thermal generators:

$$P_i^m \leq P_{i,t} \leq P_i^M \qquad i = 1, \ldots, n_g \quad t = 1, \ldots, n_t \quad (4)$$

where P_i^M, P_i^m are respectively the maximum and minimum power output of generator i.

- Maximum up and down ramps of thermal units:

$$-DR_i \leq P_{i,t} - P_{i,t-1} \leq UR_i \qquad i = 1, \ldots, n_g \qquad t = 1, \ldots, n_t \quad (5)$$

where UR_i, DR_i are respectively the maximum up and down ramp of thermal unit i.

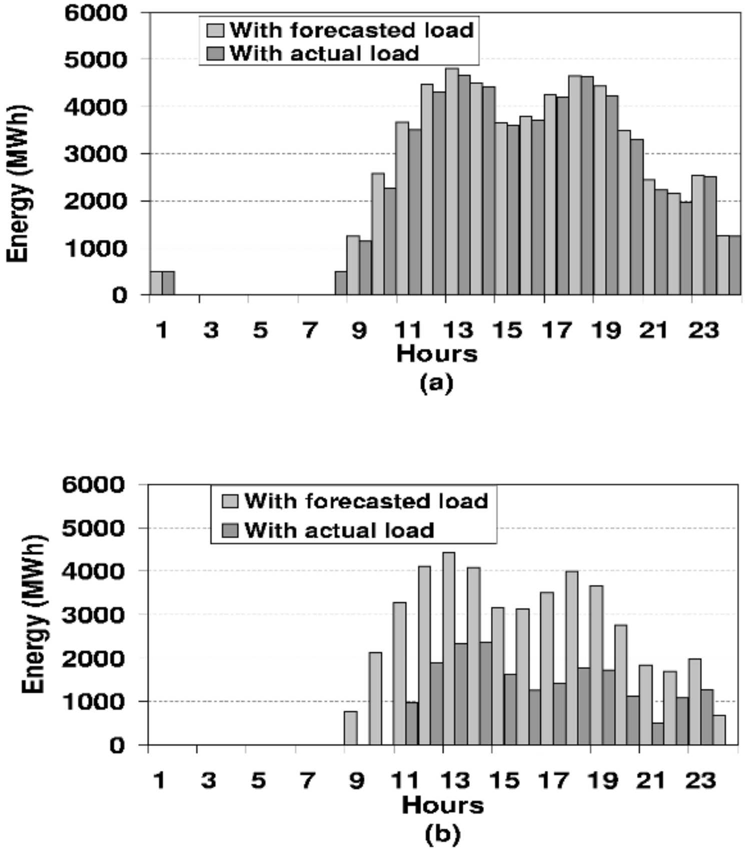

Fig. 6. Optimal scheduling of gas-turbine units for **a)** July 17th, **b)** August 6th.

- Power balance constraints:

$$\sum_{i=1}^{n_g} P_{i,t} \cdot U_{i,t} = D_t \qquad t = 1, ..., n_t \tag{6}$$

  where D_t is the system load forecasting at hour t that must be satisfied by thermal units.
- Spinning reserve constraints:

$$\sum_{i=1}^{n_g} P_i^M \cdot U_{i,t} \geq D_t + R_t \qquad t = 1, ..., n_t \tag{7}$$

  where R_t is the spinning reserve requirement at hour t.

The above model is solved by using a combined Interior Point (IP) optimization technique and a Genetic Algorithm (GA) [12]. The GA is used to compute the optimal on/off status of thermal units, while the IP module deals with the optimal solution of the short term economic dispatch, given the on/off status of the units.

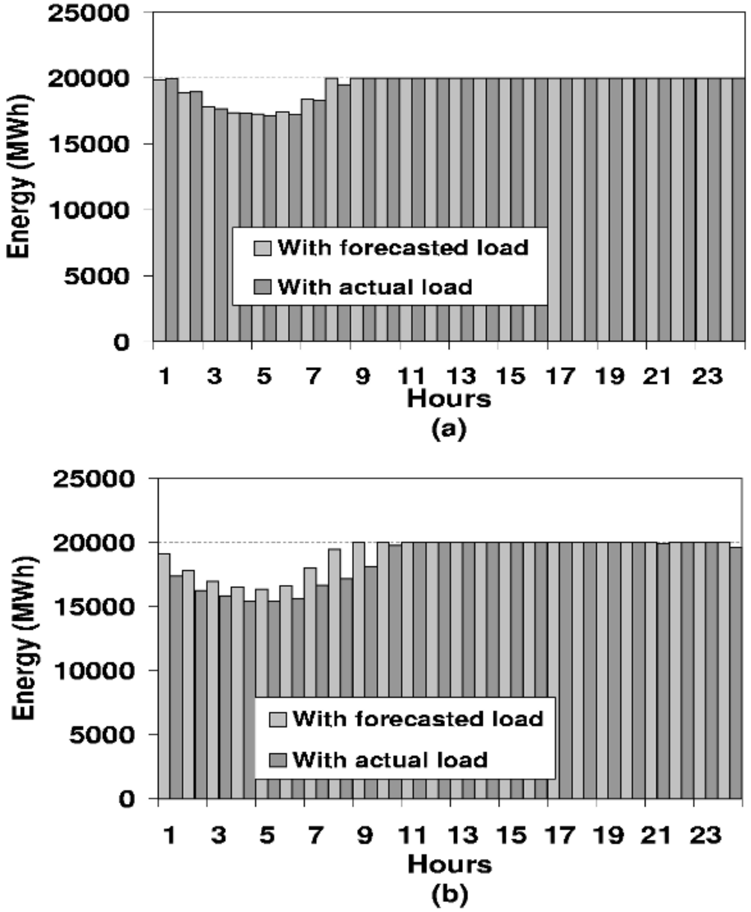

Fig. 7. Optimal scheduling of coal-fired units for **a)** July 17th, **b)** August 6th.

4 Test Results

The optimization model described in the former section is used in conjunction with the kNN-based forecasted load to assess the hourly scheduling of a test generation system comprising two generation technologies: conventional coal-fired generators and gas-turbine generators. Both generation technologies are modeled by equivalent generators with the corresponding technical characteristics.

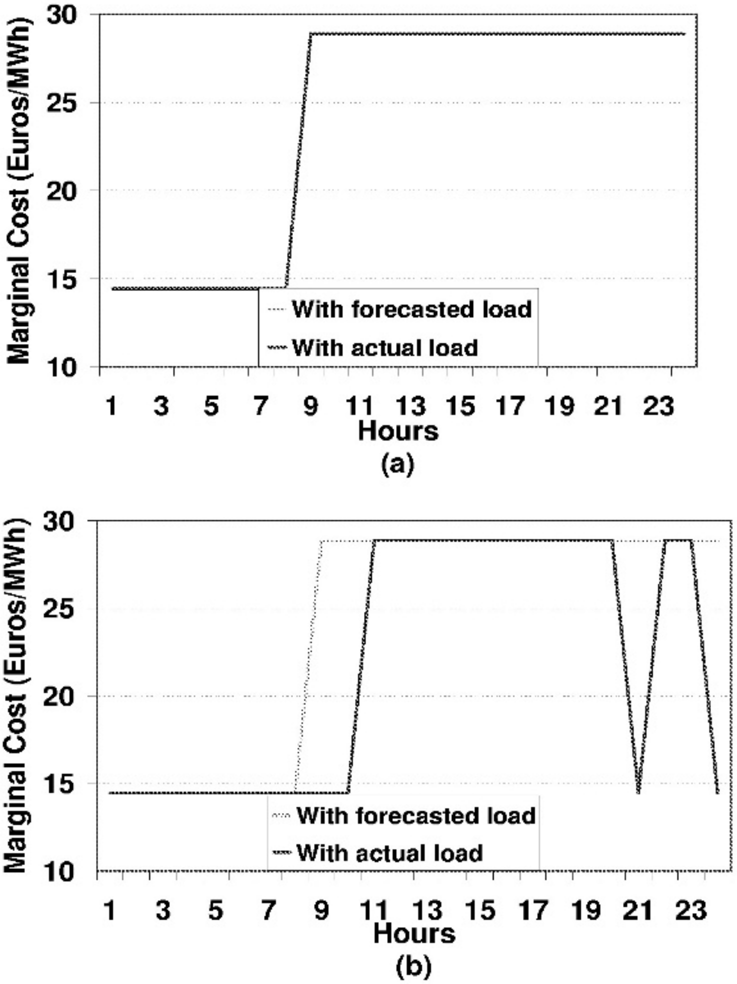

Fig. 8. Marginal price during the scheduling period for **a)** July 17th, **b)** August 6th.

The scheduling horizon embraces 24 hours. Coal-fired units take several hours to fully start and, consequently, this unit usually works at rated power. Besides, gas-turbine units are quite fast in response, and can be used to satisfy the demand at peak hours.

Figure 6a and 6b present the optimal scheduling of the gas-turbine generators attained with forecasted load and that obtained if exact load were available the day before for two selected days: July 17th and August 6th, respectively.

As discussed in the first part of the paper, these days correspond to the smallest and largest average relative errors. Note that the values of the generated power and the status on/off with forecasted and real load are very similar on July 17th, but rather different on August 6th.

Figure 7a and 7b compare the optimal scheduling of coal-fired generators obtained with forecasted load with those that would have been obtained if the actual load had been known in advance for July 17th and August 6th, respectively. As expected, coal-fired units always generate the maximum power except during the valley hours. Notice that the energy production scheduling with forecasted and actual load is almost the same on July 17th and August 6th. Thus coal-fired generators have not influence on the difference between the real total cost (when the actual load has been considered) and the approximate total cost (when the forecasted load has been considered).

Figure 8a and 8b show the optimal hourly energy cost of the generation system obtained with forecasted load and the actual load for July 17th and August 6th, respectively. It can be observed that the marginal cost with forecasted and actual load is almost the same on July 17th, with an increase on the total daily cost of only a 0.7%. However, on August 6th, the marginal cost obtained with forecasted load is larger than that obtained with the actual load at hours 9am, 10am, 9pm and 24 pm due to the gas-turbine units being started-up at these hours when the forecasted load is considered and shut-down if the actual load were used. A total cost increase of a 13.4% can be attributed to forecasting errors. It can be observed how relevant the forecasting errors are, as far as the cost of a realistic generation scheduling is concerned.

5 Conclusions

This paper addresses the influence of the accuracy of a kNN-based hourly load forecasting algorithm on the energy scheduling of a generation system. First, a kNN classification technique is proposed for load forecasting and the hourly load forecasting errors are compared with those obtained results by using a M5'. Secondly, kNN-based forecasted load profiles have been used to compute the optimal energy scheduling of a real system, and the influence of forecasting errors on the generation scheduling and the expected cost increase have been presented. The proposed algorithm based on kNN reveals much lower forecasting errors for the energy demand that the M5'. This fact is due to the M5' builds linear models and the energy demand time series is mainly nonlinear.

Acknowledgments. The authors would like to acknowledge the financial support of the Spanish Government under projects ACC-1021-TIC-2002, DPI2001-2612 and TIC2001-1143-C03-02.

References

1. A. D. Papalexopoulos and T. C. Hesterberg: A Regression-Based Approach to Short-Term System Load Forecasting. IEEE Trans. on Power System, Vol. 5, pp. 1535–1547. 1990.

2. F. J. Nogales, J. Contreras, A. J. Conejo and R. Spínola: Forecasting Next-Day Electricity Prices by Time Series Models. IEEE Trans. on Power System, Vol. 17, pp. 342–348. 2002.
3. A. S. Alfuhaid and M. A. El-Sayed: Cascaded Artificial Neural Network for Short-Term Load Forecasting. IEEE Trans. on Power System, Vol. 12, pp. 1524–1529. 1997.
4. J. Riquelme, J.L. Martínez, A. Gómez and D. Cros Goma: Load Pattern Recognition and Load Forecasting by Artificial Neural Networks. International Journal of Power and Energy Systems, Vol. 22, pp. 74–79. 2002.
5. R. Lamedica, A. Prudenzi, M. Sforna, M. Caciotta, V. Orsolini Cencellli: A Neural Network Based Technique for Short-Term Forecasting of Anomalous Load Periods. IEEE Transaction on Power Systems, Vol. 11, pp. 1749–1756. 1996.
6. A. Troncoso Lora, J. C. Riquelme Santos, J. M. Riquelme Santos, J. L. Martínez Ramos, A. Gómez Expósito: Electricity Market Price Forecasting: Neural Networks versus Weighted-Distance k Nearest Neighbours. DEXA Database Expert Systems and Applications, Aix Provence, 2002.
7. A. Troncoso Lora, J. M. Riquelme Santos, J. C. Riquelme Santos, A. Gómez Expósito, J. L. Martínez Ramos: Forecasting Next-Day Electricity Prices based on k Weighted Nearest Neighbours and Dynamic Regression. IDEAL Intelligent Data Engineering Autamitized Learning, Manchester, 2001.
8. B.V. Dasarathy : Nearest neighbour (NN) Norms: NN pattern classification techniques. IEEE Computer Society Press, 1991.
9. R. D. Short, K. Fukunaga: The Optimal Distance Measure for Nearest Neighbour Classification. IEEE Transaction on Information Theory, 1981.
10. K. Fukunaga, T. E. Flick: An Optimal Global Nearest Neighbour Metric. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1984.
11. A.J. Wood and B.F. Wollenberg, *Power Generation, Operation and Control.* John Wiley & Sons, 1996.
12. J. L. Martínez Ramos, A. Troncoso Lora, J. Riquelme Santos, A. Gómez Expósito: Short Term Hydro-thermal Coordination Based on Interior Point Nonlinear Programming and Genetic Algorithms. IEEE Porto Power Tewch Conference, 2001.
13. G. Holmes, M. Hall, E. Frank: Generating Rule Sets from Model Trees. Australian Joint Conference on Artificial Intelligence, 1999.
14. J. R. Quinlan: Learning with Continuous Classes. Australian Joint Conference on Artificial Intelligence, 1992.
15. Y. Wang, I. H. Witten: Induction of Model Trees for Predicting Continuous Classes. European Conference on Machine Learning, 1997.

Appendix: Continuous Class Prediction

This appendix briefly presents the M5' [13,14] learning algorithm used in the present work to establish a comparison between the results of the application of this algorithm and the proposed kNN classification technique on the next day energy demand forecasting problem.

In machine learning, it is important to present results that can be easily interpreted. Decision trees based on If-Then rules are one of the most popular description languages used in machine learning.

Basically, M5' builds a tree-based piecewise linear model. Model trees are decision trees with linear models at the leaf nodes. Thus, this method obtains ordered sets of If-Then rules for time series prediction that produces understandable models.

In general, the rule learning procedure for classification techniques can be stated in two steps:

1. Initially, rules are induced.
2. Rules are improved solving a global optimization problem.

The M5' algorithm builds a tree by splitting the data based on the values of predictive attributes. Once the tree has been constructed, this method computes a linear model for each node. Then the tree is pruned from the leaves while the estimated error decreases. The error for each node is the mean of the absolute difference between the predicted and the actual value of each example of the training set that reaches the node. This mean is multiplied by a weight that takes into account the number of examples that reach the node. The process is repeated until all examples are covered by one o more rules.

The M5' described is implemented in the WEKA Library [15].