2008 Eighth IEEE International Conference on Data Mining

LBF: A Labeled-Based Forecasting Algorithm and its Application to Electricity Price Time Series

Francisco Martínez–Álvarez
Area of Computer Science
Pablo de Olavide University
fmaralv@upo.es

Alicia Troncoso
Area of Computer Science
Pablo de Olavide University
ali@upo.es

José C. Riquelme
Department of Computer Science
University of Seville
riquelme@lsi.us.es

Jesús S. Aguilar–Ruiz
Area of Computer Science
Pablo de Olavide University
aguilar@upo.es

***Abstract*—A new approach is presented in this work with the aim of predicting time series behaviors. A previous labeling of the samples is obtained utilizing clustering techniques and the forecasting is applied using the information provided by the clustering. Thus, the whole data set is discretized with the labels assigned to each data point and the main novelty is that only these labels are used to predict the future behavior of the time series, avoiding using the real values of the time series until the process ends. The results returned by the algorithm, however, are not labels but the nominal value of the point that is required to be predicted. The algorithm based on labeled (LBF) has been tested in several energy-related time series and a notable improvement in the prediction has been achieved.**

I. Introduction

Time series analysis is often associated with the discovery and use of patterns –such as periodicity, seasonality or cycles– and prediction of future values, specifically termed *forecasting* in the time series context.

Therefore one may wonder what are the differences between traditional time series analysis and data mining on time series. One key difference is the large number of series involved in time series data mining. Due to the sheer amount of data involved, a highly automated modeling approach becomes indispensable in such applications. As shown in Box and Jenkins [2] and a vast volume of time series literature, traditional time series analysis and modeling tend to be based on non-automatic and trial-and-error approaches. When a large number of time series are involved, development of time series models using a non-automatic approach becomes impractical. In addition to automatic model building, discovery of knowledge associated with events known or unknown a priori can provide valuable information toward the success of a business operation.

In this paper, real-life cases are addressed in order to show the need for and the benefits of data mining on time series. The recent deregulation in electricity markets has turned this sector into a free competence scenario in which producers, investors, traders or qualified buyers can participate. Thus, the price of the electricity is determined on the basis of this buying/selling system. As a consequence, a will of obtaining optimized bidding strategies has arisen in the electricity-producer companies [22], needing both insight into future electricity prices and assessment of the risk of trusting in predicted prices.

The uncertainty of the evolution of the electricity prices is a widely studied topic. However, forecasting electricity prices is a specially difficult task because unlike demand time series, prices time series present nonconstant mean and variance and significant outliers. In that way, forecasting techniques are acquiring significant importance. Actually, several forecasting techniques have already been used to predict miscellaneous electricity time series.

Indeed, Conejo et al. [5] used the wavelet transform and ARIMA models [2] to predict the day-ahead electricity price. The authors decompose the available historical price series in four constitutive series by using the wavelet transform [12]. Then, specific ARIMA models are applied to three of these series (the fourth one is the main component of the transform) and the results are anti-transformed, providing the final forecasting. In [11] two new mixed models were proposed to obtain the forecasts of the prices in two different prediction horizons. The first one, forecasts electricity prices for each of the 24 hours of the next day using ARIMA models. They used the model estimated for one hour with the whole previous weeks to make a prediction. The second model computes the predictions for either working days or weekends using Bayesian Information Criteria.

Equally noticeable was the approach proposed by García et al. [10] in which a forecasting technique based on a GARCH model [8] was presented. Hence, this paper focuses on day-ahead forecast of electricity prices with high volatility periods. First, they apply a logarithmic transformation in order to smooth the volatility effect. Secondly, the observation of the autocorrelation helped the authors to make the selection of a specific model that deals with the seasonality of the data and the time-varying nature of volatility.

Recently, a mixing of Artificial Neural Networks [18] and Fuzzy Logic [14] was proposed in [1]. With reference to the neural network presented, it had an inter-layer and a feed-forward architecture consisting of three layers, where the hidden nodes of the proposed Fuzzy Neural Network perform the fuzzification process. Another neural network approach can be found in [3] where multiple combinations were evaluated. These combinations included networks with different number

1550-4786/08 $25.00 © 2008 IEEE
DOI 10.1109/ICDM.2008.129

of hidden layers, different number of units in each layer and different types of transfer functions.

An adaptive non-parametric regression approach was handled in [25]. The multivariate adaptive regression splines technique [9] is basically an adaptive piece-wise regression approach. This method had already been used in other predictives and data mining applications. However, it is in this work where this technique has been firstly and successfully used for electricity market price forecasting purposes.

A modification of the Nearest Neighbors methodology [7] is proposed in [23]. To be precise, the approach weights the nearest neighbors so that the forecasting is improved.

The occurrence of spike prices (price that is significantly higher than its expected value) is an usual peculiarity associated to price time series. With the aim of dealing with this feature, the authors in [26] proposed a data mining framework based on both support-vector machines [6] (SVM) and probability classifier.

Li et al. proposed a forecasting system immersed in a grid environment in [15]. In this paper, a fuzzy inference system, adopted due to its transparency and interpretability, and time series methods are proposed for day-ahead electricity price forecasting.

Despite the variety of data mining techniques used in order to perform the prediction of the prices, none of them are based only on the labels generated by using clustering techniques. The novel and main contribution of this paper is, therefore, a new algorithm that only uses these labels to predict the future behavior of a time series, avoiding using the real values of the time series until the process ends. Hence, this work tackle the problem in a framework based on non-supervised learning, which will enhance the prices prediction accuracy, providing a new procedure to perform forecasts. Moreover, all the data sets analyzed are available on-line in order to facilitate the comparison of the results obtained.

The rest of the paper is organized as follows. Section 2 introduces the proposed methodology and the LBF algorithm is presented, providing a method to apply in time series of any nature. Section 3 shows the results obtained by the LBF approach in electric energy markets of Spain, Australia and New York for the whole year 2006, giving a measure of the quality of them. In Section IV comparisons between the proposed method and other techniques are shown. Finally, Section V expounds the conclusions achieved and gives clues for future works.

II. The proposed methodology

The proposed methodology is divided in two phases clearly differentiated. In a first step, a clustering technique is performed and, secondly, the phase of forecasting is applied using the information provided by this clustering. The LBF forecasting algorithm is focused on predicting samples framed in a time series, either one-dimensional or multi-dimensional, previously labeled with clustering techniques. By using this strategy, two advantages are enjoyed. From one side, it reduces the dimensionality of the data with the resulting time processing decrease. As soon as the clustering is applied, the algorithm only processes the number of cluster –the label– assigned to the samples, ignoring if they had more than one feature. On the other hand, the complexity of the algorithm is drastically reduced insofar as the computation process is directly proportional to the dimensionality of the data.

The LBF method allows predicting more than one sample because it is implemented with a close loop that feeds the sample-ahead prediction back in the data set, in order to predict the following sample. This feature is especially useful when the horizon of prediction has to cover various samples. Figure 1 shows the basic idea behind the proposed methodology.

A. Data normalization

The first task to be completed is the normalization of the data. It can be assumed that the prices increase all along the year following a tendency in accordance with the intra-annual inflation. That is, the original trend is suppressed from the initial data; otherwise it could muddle up the results. The transformation applied is:

$$p_j \leftarrow \frac{p_j}{\frac{1}{N}\sum_{i=1}^{N} p_j} \tag{1}$$

where p_j is the price of the $j-th$ hour of the day and N the number of samples considered per day. In this case, $N = 24$ since each sample represents one hour of the day.

B. Clustering technique

At this point the data has already been conveniently preprocessed and cleared. Clustering techniques are, now, going to be applied to label time series.

Given the data base of hourly prices the clustering problem consists of identifying K groups or clusters such that the prices curves of the days belonging to a cluster are similar between them and disimilar to the prices curves of the days belonging to other clusters, according to a distance measure.

As a consequence, the dimensionality of the data base is drastically reduced from its initial 24 features (equivalent to the 24 hours of the day) to only one dimension (the label of the cluster which the day belongs). This effect can be observed in Figure 2.

To achieved this challenge, two questions have to be answered: which clustering technique has to be chosen? and, if it is appropriate, how many clusters has to be created?

These two topics has widely been discussed in the literature [24]. Nevertheless, it seems that there is not an unique answer because it depends on many subtle factors.

Hard or fuzzy clustering are the two main branches of non-supervised classification techniques that can be used. Once the data are prepared, a clustering technique is applied in order to label each daily electricity price curve. The discussion of choosing one technique or another can be found in [17], in which the well-known K-means algorithm was the optimum method to classify this kind of data set.

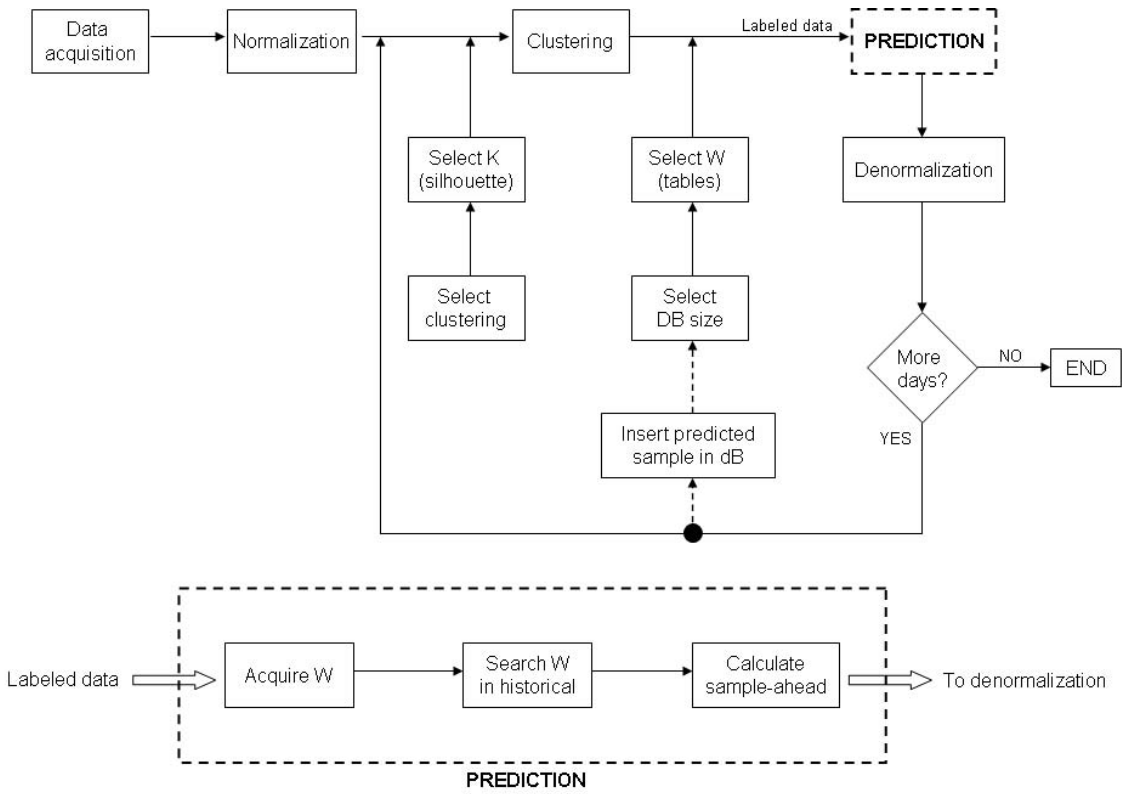

Fig. 1. Illustration of the proposed methodology. The prediction stage is further detailed.

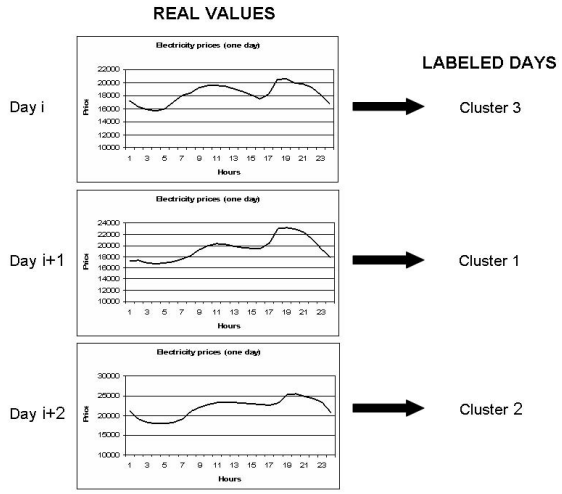

Fig. 2. Assigning one label for each day.

However, the K-means algorithm requires that the user provides the number of clusters to be created. For this reason, the *silhouette function* [13] was used to decide in how many groups the original data set has to be split. The *silhouette function* provides a measure of the quality of the separation between the clusters obtained by using the K-means algorithm.

In an object i belonging to the cluster C_k, the average dissimilarity of i to all other objects of C_k is denoted by $c_k(i)$. Analogously, in cluster C_m, the average dissimilarity of i to all objects of C_m is called $dis(i, C_m)$. After computing $dis(i, C_m)$ for all clusters $C_m \neq C_k$, the smallest one is selected as follows,

$$c_m(i) = min\{dis(i, C_m)\}, \forall m \text{ such that } C_m \neq C_k. \quad (2)$$

This value represents the dissimilarity of the object i to its neighbor cluster. Thus, the silhouette values, $silh(i)$ are given by the following equation:

$$silh(i) = \frac{c_k(i) - c_m(i)}{max\{c_k(i), c_m(i)\}} \quad (3)$$

The $silh(i)$ can vary between -1 and $+1$, where $+1$ denotes clear cluster separation and -1 marks points with questionable cluster assignment. If cluster C_k is a singleton, then $silh(i)$ is not defined and the most neutral choice is to set $silh(i) = 0$. The objective function is the average of $silh(i)$ over the number of objects to be classified, and the best clustering is reached when the above mentioned function is maximized.

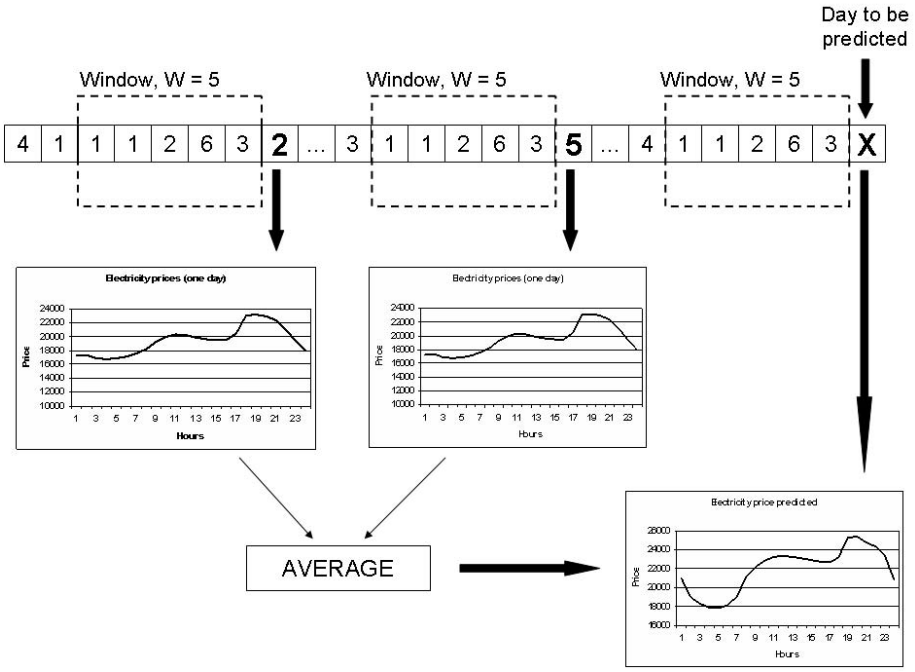

Fig. 3. LBF algorithm.

C. The LBF algorithm

Given the hourly prices recorded in the past, up to day d, the forecasting problem aims at predicting the 24 hourly prices corresponding to day d+1.

Let $P_i \in \mathbf{R}^{24}$ be a vector composed of the 24 hourly energy prices corresponding to a certain day i

$$P_i = [p_1, p_2, \ldots, p_{24}]\,. \tag{4}$$

Let L_i be the label of the prices of the day i obtained as a previous step to the forecasting by using a clustering technique. Let S_W^i the subsequence of labels of the prices of the W consecutive days, from day i backward, as follows,

$$S_W^i = [L_{i-W+1}, L_{i-W+2}, \ldots, L_{i-1}, L_i] \tag{5}$$

where the length of the window, W, is a parameter to be determined.

The LBF algorithm first searches the subsequences of labels which are exactly equals to S_W^d in the data base, providing the equal subsequences set, ES, defined by this equation,

$$ES = \left\{\text{set of indexes } j \text{ such that } S_W^j = S_W^d\right\} \tag{6}$$

In case of not finding any subsequence in data base equal to S_W^d, the procedure searches the subsequences of labels which are exactly equals to S_{W-1}^d. That is, the length of the window composed of the subsequence of labels is decreased.

According to the LBF approach, the 24 hourly prices of day $d+1$ are predicted by averaged the prices of the days succeeding those in ES. That is,

$$P_{d+1} = \frac{1}{\text{size}(ES)} \cdot \sum_{j \in ES} P_{j+1} \tag{7}$$

where size(ES) is the number of elements belonging to the set ES. Afterwards, LBF algorithm outputs need to be de-normalized to generate the desired forecasted values.

This procedure is detailed in Figure 3.

In case of a long-term prediction, in which more than one forecasted sample is required, the following tasks have to be carried out. First of all, the real values of the predicted sample are linked to the whole data set. Second, the clustering process is repeated with the enlarged data set and, finally, the window size is re-calculated and the prediction step is performed (to see Figure 1).

D. Selecting the size of the window

The previous clustering generates a sequence of labels associated to every day. Now, a subsequence of labels is taken into consideration for further steps; concretely, if the day $d+1$ has to be predicted, the sequence of labels $S_W^d = [L_{d-W+1}, L_{d-W+2}, \ldots, L_{d-1}, L_d]$ is extracted from the data set and it is used as a pattern of search, where W is the length of this subsequence or window.

This stage is, perhaps, the most critical of the whole process insofar as a wrong value for W may affect deeply in the rest of the forecasting. The selection of W depends on the case under study but it can be systematically tuned. Thus, it is compulsory to perform a training phase to find an adequate value for W before applying the LBF approach. This step is illustrated in Figure 4.

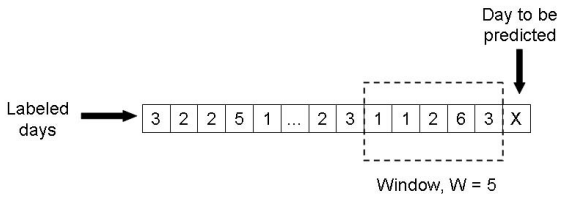

Fig. 4. Optimal window length.

The optimal number of labels contained in the window that will be used as a pattern of search to find all equal subsequences of labels in data base (parameter W) is determined minimizing the forecasting error when the LBF method is applied to the training set.

Mathematically, this means to find the value of W that minimizes the following function,

$$\sum_{d \in TS} |\widehat{P}_{d+1} - P_{d+1}| \tag{8}$$

where $\widehat{P}_{d+1}$ are forecasted prices for day $d+1$, according to the LBF method, P_{d+1} are actual recorded prices and TS refers to the training set. Notice that, according to (7), P_{d+1} is an implicit function of the discrete variable W. Hence, the application of standard mathematical programming methods is not possible when searching for W. In practice, W is assigned successive integer numbers ($W = 2, 3, \ldots$) until a local minimum is found.

III. Results

The first goal to be fulfilled is to find those time series whose prediction have relevance. This work is focused on predicting electricity price time series including clustering techniques as a previous task. In order to prove that the algorithm works properly over any kind of data set, several public electricity prices time series have been considered. To be precise, the methodology described above has been applied to the electricity prices of Spanish [19], Australian [16] and New York [20] markets.

This section is structured as follows. First, the LBF has to be trained in order to produce accurate predictions and, for this reason, the election of both W and K is discussed here. Second, the accuracy of the predictions has to be somehow validated. Thus, some quality parameters are presented. Third, the prediction of the year 2006 is provided.

A. Training the LBF

In this subsection the number of clusters to be generated, K, as well as the length of the window, W, that has to be searched all along the time series, is presented. This step has to be repeated every time the kind of the time series changes.

First of all, the number of clusters K has to be chosen and, for this purpose, a subsequence of twelve months is considered. From all these twelve months, eleven are used for training the algorithm and the 12-th is utilized in order to make predictions. Table I summarizes the periods used in the three time series analyzed.

TABLE I
PERIODS USED TO CALCULATE PARAMETERS K AND W.

Market	Training period	Evaluated on
Spanish Market	Dec 2001 - Oct 2002	Nov 2002
Australian Market	May 2002 - Mar 2003	Apr 2003
New York Market	Feb 2004 - Dec 2004	Jan 2005

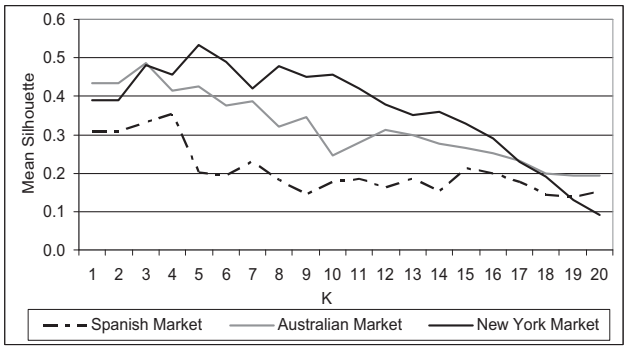

Fig. 5. The mean value of *silhouette* when varying K for the three markets.

According to the methodology proposed in [17], the silhouette function is applied to these three time series. Figure 5 shows the variation of the mean silhouette value with relation to the number of clusters, K. When the curves reach their higher values, it can be stated that the corresponding K value (X axis) is the one that generates the best clusters possible, that is, the intra-cluster distance is minimized and the inter-cluster is maximized. As it can be appreciated, the number of clusters selected were $K = 4$, $K = 3$ and $K = 5$ for the Spanish, Australian and New York markets, respectively. The Figures 6, 7 and 8 illustrate the silhouette curves obtained when the Spanish, Australian and New York Markets are evaluated respectively with the above mentioned values of K.

As the number of clusters is already decided, the next step consists in selecting the optimal length of the window W.

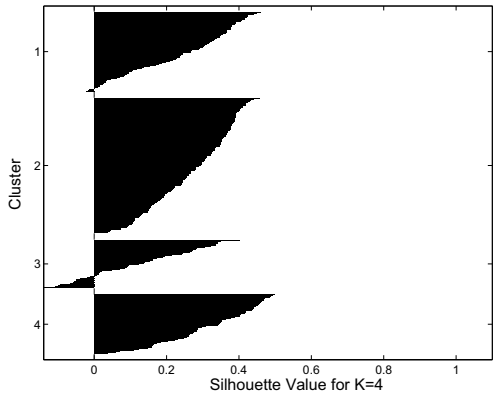

Fig. 6. Silhouette function when $K = 4$ in the Spanish Market.

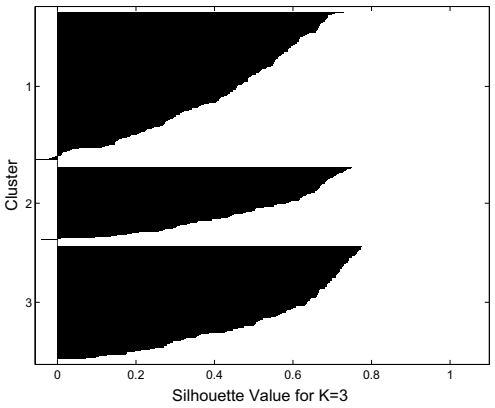

Fig. 7. Silhouette function when $K = 3$ in the Australian Market.

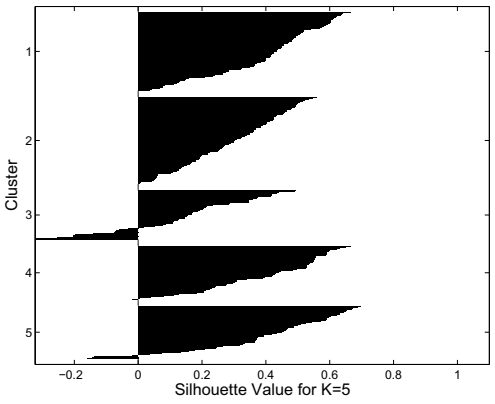

Fig. 8. Silhouette function when $K = 5$ in the New York Market.

Thus, this step is focused on finding the W that obtains the minimum prediction error.

Therefore, it is required to evaluate the performance of the LBF algorithm when W varies. Table II shows how the prediction error varies in accordance with the number samples considered in the window. A 100% error means that such a long sequence was not found when K clusters were considered in the training set. Finally, the W that allows a lower prediction error is the value chosen for further forecasting on real data. From the observation of the Table II, it can be concluded that the lengths of the windows that have to be used are $W = 5$, $W = 6$ and $W = 3$ for the Spanish, Australian and New York electricity markets respectively. The results of training the LBF are summarized on Table III.

TABLE III
NUMBER OF CLUSTERS K AND LENGTH OF THE WINDOW W PARAMETERS FOR THE THREE ELECTRICITY PRICE TIME SERIES.

Electricity Price Market	K	W
Spanish	4	5
Australian	3	6
New York	5	3

B. Parameters of quality.

To evaluate the accuracy of the LBF approach in forecasting time series different criteria could be used. However, the most relevant parameters which have to be taken into consideration are:

- Mean relative error to $\bar{p}$ (MRE).

$$MRE = 100 \cdot \frac{1}{N} \sum_{h=1}^{N} \frac{\mid \hat{p}_h - p_h \mid}{\bar{p}} \tag{9}$$

where

$$\bar{p} = \frac{1}{N} \sum_{h=1}^{N} p_h \tag{10}$$

$\hat{p}_h$ and p_h are the predicted and current electricity prices at hour h respectively, $\bar{p}$ is the mean price for the period of interest (a day or a week in this work) and N is the number of predicted hours. Note that, the mean price is used in the denominator of (9) to avoid the effect of prices close to zero.

- Mean squared error (MSE)

$$MSE = \frac{1}{N} \sum_{h=1}^{N} (\hat{p}_h - p_h)^2 \tag{11}$$

- Standard deviation of relative error (σ_{MRE}).

$$\sigma_{MRE} = \sqrt{\frac{1}{N} \sum_{h=1}^{N} (e_h - \bar{e})^2} \tag{12}$$

where

$$e_h = \frac{\hat{p}_h - p_h}{\bar{p}} \tag{13}$$

and

$$\bar{e} = \frac{1}{N} \sum_{h=1}^{N} e_h \tag{14}$$

C. Results of forecasting year 2006

In this subsection the results obtained when the LBF algorithm was applied into the three different markets is provided. Precisely, Tables IV, V and VI show the MRE, MSE and σ_{MRE} produced in the Spanish, Australian and New York markets when the year 2006 was taken into consideration.

Figure 9 illustrates the best prediction curve obtained for the Spanish market in the year 2006 in cents of Euro per KWHr (cE/KWHr). It took place for 23^{rd} June and its MRE was 3.10%. On the contrary, Figure 10 references the worst prediction. It took place the 8^{th} May and its MRE was 9.39%.

It is important to remark that the Australian market shows their information structured in different areas. Thus the National Electricity Market in Australia is comprised of five jurisdictions: Queensland, New South Wales, Victoria, Tasmania and South Australia. The results in Table V refers to the Queensland Market.

Figure 11 illustrates the best prediction curve obtained for the Australian market in the year 2006 in dolars per MWHr ($/MWHr). It took place for 12^{th} May and its MRE

TABLE II
PREDICTION ERROR PERFORMED BY THE LBF ALGORITHM ON THE TEST SETS.

Electricity Price Market	W=1	W=2	W=3	W=4	W=5	W=6	W=7	W=8	W=9	W=10
Spanish Market ($K = 4$)	10.32%	8.44%	8.21%	4.39%	**2.23%**	2.89%	100%	100%	100%	100%
Australian Market ($K = 3$)	9.58%	7.91%	6.26%	6.17%	7.33%	**5.81%**	6.04%	9.12%	100%	100%
New York Market ($K = 5$)	7.09%	5.98%	**3.27%**	6.98%	4.45%	13.20%	10.31%	100%	100%	100%

TABLE IV
PREDICTION ACCURACY OF THE LBF ALGORITHM FOR THE YEAR 2006 IN THE SPANISH ELECTRICITY MARKET.

Month	MRE	MSE	σ_{MRE}
January	7.26%	0.34	0.25
February	4.93%	0.45	0.19
March	5.88%	0.33	0.22
April	3.62%	0.37	0.18
May	8.11%	0.45	0.21
June	3.76%	0.21	0.24
July	4.30%	0.35	0.23
August	5.37%	0.37	0.34
September	6.41%	0.37	0.31
October	7.89%	0.41	0.29
November	8.30%	0.46	0.40
December	8.02%	0.43	0.36
Average	6.15%	0.38	0.27

TABLE V
PREDICTION ACCURACY OF THE LBF ALGORITHM FOR THE YEAR 2006 IN THE AUSTRALIA'S NATIONAL ELECTRICITY MARKET.

Month	MRE	MSE	σ_{MRE}
January	5.58%	2.31	1.34
February	8.59%	6.42	3.24
March	7.84%	5.87	2.98
April	9.92%	6.27	3.90
May	12.85%	9.12	4.03
June	22.04%	24.54	12.34
July	17.11%	22.76	10.58
August	11.71%	8.34	5.08
September	8.23%	6.23	2.45
October	7.66%	5.01	2.89
November	6.76%	4.81	1.94
December	6.42%	3.82	2.01
Average	10.39%	8.79	4.40

TABLE VI
PREDICTION ACCURACY OF THE LBF ALGORITHM FOR THE YEAR 2006 IN THE NEW YORK INDEPENDENT SYSTEM OPERATOR.

Month	MRE	MSE	σ_{MRE}
January	4.45%	5.01	4.32
February	5.53%	4.56	2.34
March	6.30%	9.04	6.42
April	4.94%	6.78	2.18
May	7.59%	12.26	4.56
June	3.34%	5.67	3.72
July	3.93%	5.89	2.86
August	5.37%	4.74	3.56
September	6.24%	8.17	3.04
October	7.43%	9.98	5.53
November	5.19%	8.34	4.44
December	6.04%	7.30	3.98
Average	5.53%	7.31	3.91

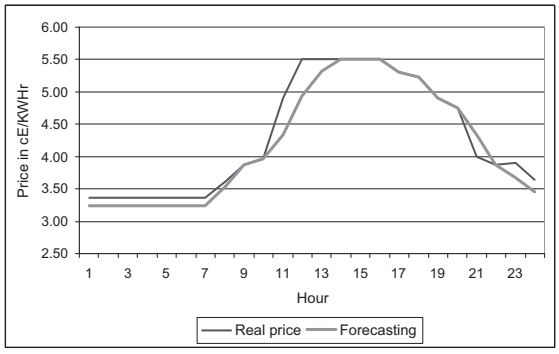

Fig. 9. Best prediction reached for the Spanish electricity prices market.

was 3.66%. On the contrary, Figure 12 references the worst prediction. It took place the 20^{th} July and its MRE was 65.60%.

Figure 13 illustrates the best prediction curve obtained for the New York market in the year 2006 in dolars per MWHr ($/MWHr). It took place for 8^{th} July and its MRE was 2.76%. On the contrary, Figure 14 references the worst prediction. It took place the 12^{th} May and its MRE was 8.89%.

IV. COMPARING THE LBF PERFORMANCE WITH OTHER TECHNIQUES

A comparison between the results obtained with the LBF method and many other approaches is provided in this section, demonstrating that LBF approach improves all existing techniques used in this area. Thus, in order to validate somehow the accuracy of the proposed algorithm, it has been applied to specific periods of time in which others authors evaluated their own approaches.

The *Spanish electricity price market* has been widely analyzed. Many authors have proved their own novel approaches in the year 2002 and, as a consequence, the literature offers multiples results in this year. The LBF algorithm is compared with the four most recently approaches published: ARIMA [5], Neural Networks [3], Mixed Models [11] and Weighted Nearest Neighbors [23]. Finally, it is also compared with the Naïve Bayes classifier [21]. As it can be appreciated in Table VII, the proposed method has improved all the MRE rates.

The authors in [11] also forecasted a week of the year 2000. The comparative MRE rates are shown in Table VIII.

The prices in the *Australia's National Electricity Market* have also been predicted in [26]. It is remarkable that this market presents an especial behavior since many spot prices are observed. Despite the authors in [26] have developed techniques based on support-vector machines in order to deal with this particular days, the LBF algorithm does not make any assumption about the nature of the days to be predicted,

TABLE VII
COMPARISON OF THE MRE PROVIDED BY LBF, ARIMA, NEURAL NETWORKS, NAÏVE, WNN AND MIXED MODELS.

Week	Naïve	Neural Networks	ARIMA	Mixed Models	WNN	**LBF**
18^{th}–24^{th} Feb 2002	7.68%	5.23%	6.32%	6.15%	6.01%	5.98%
20^{th}–26^{th} May 2002	7.27%	6.36%	6.36%	4.46%	5.99%	4.51%
19^{th}–25^{th} Aug 2002	27.30%	11.40%	13.39%	14.90%	11.23%	9.11%
18^{th}–24^{th} Nov 2002	19.98%	13.65%	13.78%	11.68%	11.59%	10.07%
Average	15.56%	9.16%	9.96%	9.30%	8.71%	7.42%

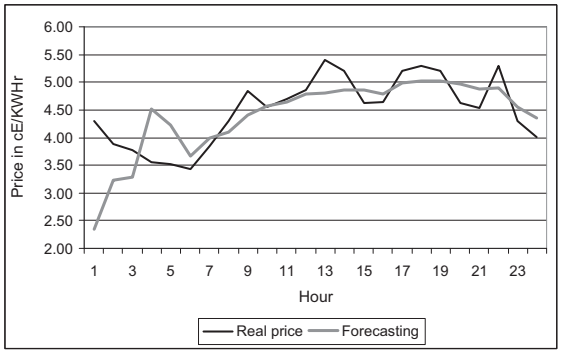

Fig. 10. Worst prediction reached for the Spanish electricity prices market.

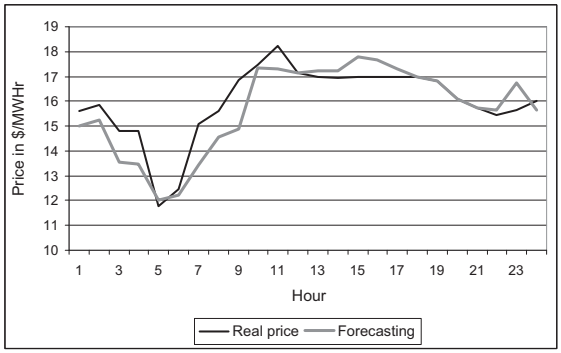

Fig. 11. Best prediction reached for the Australian electricity prices market.

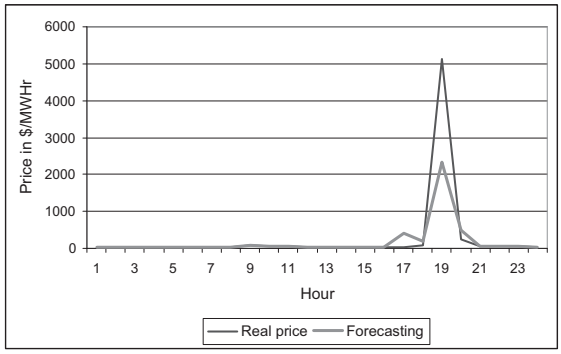

Fig. 12. Worst prediction reached for the Australian electricity prices market.

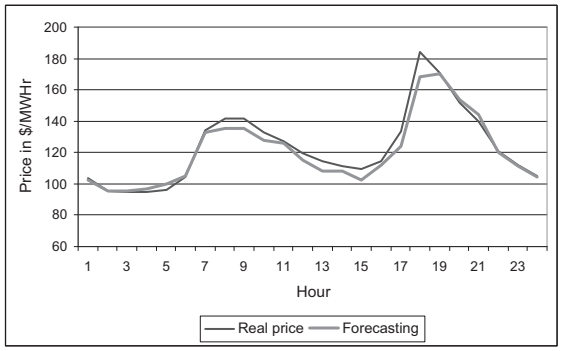

Fig. 13. Best prediction reached for the New York electricity prices market.

insofar it uses unsupervised learning and, consequently, no a priori information is known about data.

The MRE supplied in Table IX are about, precisely, these days with spike prices in the year 2004.

As for the *New York electricity price time series,* the authors in [4] compared some forecasting algorithms with their own approach. They applied manifold-based dimensionality reduction to electricity price curve modeling. Hence, they demonstrated that it exists a low-dimensional manifold representation for the day-ahead price curve in the New York electricity market.

The results in Table X stand for the MRE of one week-ahead

TABLE VIII
MRE FOR AUGUST 25^{th}–31^{st} 2000 IN THE SPANISH MARKET.

Day	ARIMA	Mixed Models	**LBF**
Day 1	4.30%	4.80%	3.74%
Day 2	7.99%	7.30%	6.91%
Day 3	4.57%	5.40%	3.45%
Day 4	10.81%	4.60%	5.21%
Day 5	6.12%	5.10%	4.48%
Day 6	17.34%	14.90%	9.63%
Day 7	6.05%	7.20%	4.81%
Average	8.17%	7.04%	5.46%

TABLE IX
MRE FOR SOME DAYS IN JUNE 2004 IN THE AUSTRALIAN MARKET.

Day (2004)	ARIMA	SVM	**LBF**
5^{th} June	32.31%	18.09%	16.72%
17^{th} June	29.09%	13.31%	8.31%
20^{th} June	33.73%	17.11%	14.23%
21^{st} June	24.18%	19.20%	18.93%
Average	29.82%	16.93%	14.55%

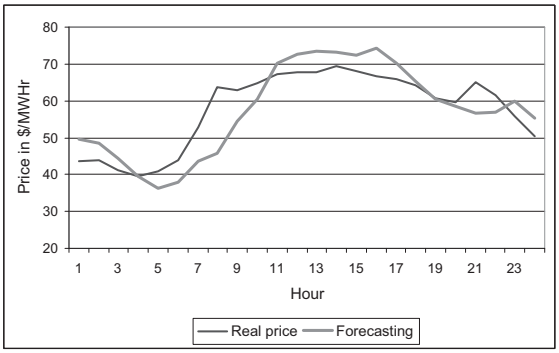

Fig. 14. Worst prediction reached for the New York electricity prices market.

TABLE X
PREDICTION ACCURACY OF THE LBF FOR THE YEAR 2005 IN THE NEW YORK INDEPENDENT SYSTEM OPERATOR.

Month	Naïve	ARIMA	STR	**LBF**
Feb 2005	15.84%	8.14%	7.37%	6.99%
Mar 2005	10.06%	5.58%	5.45%	6.02%
Apr 2005	12.39%	6.11%	6.58%	6.12%
May 2005	5.83%	7.28%	6.06%	4.83%
Jun 2005	31.78%	9.67%	9.72%	5.37%
Jul 2005	17.49%	7.48%	7.61%	8.04%
Aug 2005	13.02%	5.98%	5.43%	3.51%
Sep 2005	14.67%	7.19%	7.48%	6.91%
Oct 2005	9.68%	6.37%	6.38%	5.68%
Nov 2005	18.74%	5.87%	6.10%	6.03%
Dec 2005	27.86%	8.52%	8.79%	7.01%
Jan 2006	15.42%	10.50%	8.25%	6.85%
Average	16.07%	7.39%	7.10%	6.11%

electricity price forecasting for each second week of the year 2005. The STR column corresponds to the results obtained by the structural model proposed in [4].

V. CONCLUSIONS

In this paper, a new forecasting algorithm has been proposed to predict real-world time series. As previous step to the prediction, a clustering technique to label 24-dimensional time series samples has been used and the main novelty lies on the using of only the labels obtained by the clustering to forecast the future behavior of the time series, avoiding using the real values of the time series until the process ends. The algorithm has been successfully applied in electricity prices time series of Spanish, Australian and New York markets, improving the results of the existing techniques nowadays.

Future work is focussed in tuning the model with a dynamical length of the window and in the relaxation of the set ES searching subsequences similar in a percentage as an alternative to exactly equal subsequences.

ACKNOWLEDGMENT

The authors would like to thank the financial support from the Spanish Ministry of Science and Technology, project TIN2007-68084-C-02, and from the Junta de Andalucía, project P07-TIC-02611.

REFERENCES

[1] N. Amjady. Day-ahead price forecasting of electricity markets by a new fuzzy neural network. *IEEE Transactions on Power Systems*, 21(2):887–896, 2006.

[2] G. E. Box and G. Jenkins. *Time Series Analysis Forecasting and Control*. Holden-Day, 1976.

[3] J. P. S. Catalao, S. J. P. S. Mariano, V. M. F. Mendes, and L. A. F. M. Ferreira. Short-term electricity prices forecasting in a competitive market: a neural network approach. *Electric Power Systems Research*, 77:1297–1304, 2007.

[4] J. Chen, S. J. Deng, and X. Huo. Electricity price curve modeling by manifold learning. *IEEE Transactions on Power Systems*, 15:723–736, 2007.

[5] A. J. Conejo, M. A. Plazas, R. Espínola, and B. Molina. Day-ahead electricity price forecasting using the wavelet transform and ARIMA models. *IEEE Transactions on Power Systems*, 20(2):1035–1042, 2005.

[6] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[7] B. V. Dasarathy. *Nearest Neighbor Pattern Classification Techniques*. IEEE Computer Society Press, 1991.

[8] R. F. Engle. Autoregressive conditional heteoskedasticity with estimates of the variance of the UK inflation. *Econometrica*, 50(4):987–1007, 1982.

[9] J. H. Friedman. Multivariate adaptive regression splines. *The Annual of Statistics*, 19(1):1–67, 1991.

[10] R. C. García, J. Contreras, M. van Akkeren, and J. B. García. A GARCH forecasting model to predict day-ahead electricity prices. *IEEE Transactions on Power Systems*, 20(2):867–874, 2005.

[11] C. García-Martos, J. Rodríguez, and M. J. Sánchez. Mixed models for short-run forecasting of electricity prices: Application for the spanish market. *IEEE Transactions on Power Systems*, 22(2):544–552, 2007.

[12] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIMAT*, 15:723–736, 1984.

[13] L. Kaufman and P. J. Rousseeuw. *Finding groups in Data: an Introduction to Cluster Analysis*. Wiley, 1990.

[14] G. J. Klier, U. H. St Clair, and B. Yuan. *Fuzzy set theory: foundations and applications*. Prentice Hall, 1997.

[15] G. Li, C. C. Liu, C. Mattson, and J. Lawarrée. Day-ahead electricity price forecasting in a grid environment. *IEEE Transactions on Power Systems*, 22(1):266–274, 2007.

[16] Australia's National Electricity Market. Available on-line. *http://www.nemmco.com.au*.

[17] F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme, and J. M. Riquelme. Partitioning-clustering techniques applied to the electricity price time series. *Lecture Notes in Computer Science*, 4881:990–999, 2007.

[18] W. S. McCulloch and W. Pitts. A logical calulus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[19] Spanish Electricity Price Market Operator. Available on-line. *http://www.omel.es*.

[20] The New York Independent System Operator. Available on-line. *http://www.nyiso.com*.

[21] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufman Publishers, 1988.

[22] M. A. Plazas, A. J. Conejo, and F. J. Prieto. Multimarket optimal bidding for a power producer. *IEEE Transactions on Power Systems*, 20(4):2041–2050, 2005.

[23] A. Troncoso, J. C. Riquelme, J. M. Riquelme, J. L. Martínez, and A. Gómez. Electricity market price forecasting based on weighted nearest neighbours techniques. *IEEE Transactions on Power Systems*, 22(3):1294–1301, 2007.

[24] R. Xu and D. C. Wunsch II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[25] H. Zareipour, K. Bhattacharya, and C. A. Cañizares. Forecasting the hourly Ontario energy price by multivariate adaptive regression splines. *IEEE Transactions on Power Systems*, 20(2):1035–1042, 2006.

[26] J. H. Zhao, Z. Y. Dong, X. Li, and K. P. Wong. A framework for electricity price spike analysis with advanced data mining methods. *IEEE Transactions on Power Systems*, 22(1):376–385, 2007.