Cluster preprocessing to improve time series forecasting

Francisco Martínez-Álvarez
Department of Computer Science, Pablo de Olavide University of Seville, Seville, Spain
E-mail: fmaralv@upo.es

Abstract. This work proposes a novel general-purpose forecasting algorithm. It first extracts patterns from time series using the information provided by certain clustering techniques, which are applied as a first step of the approach. Moreover, the occurrence of data with especially unexpected values (outliers) is also addressed in this work. To deal with these outliers, a new hybrid methodology has been proposed, by inserting and adapting an existing approach based on the discovery of frequent episodes in sequences in the general scheme of prediction.

Keywords: Time series, forecasting, clustering, outliers

1. Introduction

This work analyzes and forecasts time series by means of data mining. In particular, the use of clustering techniques is crucial in the proposed forecasting algorithm and a thorough analysis on its suitability can be found in the original research work [1].

The first relevant contribution is the application of clustering techniques to diverse time series to discover meaningful patterns to be used in subsequent forecasting processes.

The second main contribution consists in the development of a time series forecasting algorithm (henceforth called PSF, Pattern Sequence Forecasting) that makes use of the patterns previously generated by clustering. Thus, a new methodology is proposed based on the patterns extracted during the aforementioned process.

Finally, this work deals with especially unexpected values or outliers. Hence, PSF has been enhanced by its combination with other existing approach based on the discovery of frequent episodes in sequences.

2. Pattern sequences to forecast time series

The PSF scheme is now described. It first discretizes the data by means of clustering techniques, assigning a cluster – or label – to each sample that forms the time series. Then, the labels preceding the sample to be predicted are selected and this sequence of labels, or pattern sequence, is searched in the historical data. Every time the pattern sequence is found, the sample after the pattern is evaluated. In particular, the prediction provided by PSF is the average values of the samples found after every matching. Note that the algorithm returns real-valued forecasts instead of labels.

An initial version of the algorithm can be found in [3]. However, the final version of PSF was introduced in [5], which also included a new methodology to automatize the labels assignment to the real-valued samples. The sensitivity of the parameters involved in the selection of the number of patterns has been also analyzed in order to study the robustness of the method. Last, the number of labels comprising the pattern sequence is also systematically determined.

With reference to the clustering technique used in PSF, several techniques have been considered. Since each one tends to discover clusters with fixed shapes, the use of K-means, Expectation–Maximization and Fuzzy C-means has been discussed to determine their adequacy for every time series. Hence, a new methodology to systematically select the number of clusters has been proposed in [5]. This strategy is based on a...
Table 1: PSF algorithm forecasting errors

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Prices 2006</th>
<th>Demand 2006</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MER (%)</td>
<td>σ_MER</td>
</tr>
<tr>
<td>Spain</td>
<td>6.15</td>
<td>0.27</td>
</tr>
<tr>
<td>New York</td>
<td>5.53</td>
<td>1.94</td>
</tr>
<tr>
<td>Australia</td>
<td>10.39</td>
<td>4.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Ozone 2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>19.02</td>
</tr>
</tbody>
</table>

majority-based vote and combines all the three aforementioned techniques.

In order to assess the PSF performance, it has been applied to seven different real-world time series (a total of 84 datasets), six of them related to electricity prices and demand markets, and another one to the tropospheric ozone. A summary of the mean error relative (MER) and its standard deviation (σ_MER) that PSF obtained, can be found in Table 1.

Moreover, its accuracy has been evaluated by comparing the forecasting results to those of other approaches found in the literature and it provided lower errors in all cases. In particular, support vector machines, discrete Wavelet transforms, artificial neural networks, dynamic regressions and ARIMA models have been applied to more than 30 datasets (different from those in Table 1), reporting a 14.32% MER, on average. By contrast, PSF obtained mean error of 7.03% on average when applied to these datasets.

On the other hand, patterns have been discovered in seismic time series for the periods of time elapsed just before an earthquake occurs [2]. These patterns revealed certain masked events that were eventually considered as earthquake precursors. Note that for the discovery of these patterns, only clustering techniques were analyzed (there is no forecasting provided in [2], just patterns that lead to large earthquakes). In particular, two seismological Spanish areas have been studied, obtaining a sensitivity of 90.00 and 79.31%, and a specificity of 82.56 and 90.38%, respectively.

Additionally, the occurrence of outliers has been studied. To be precise, an existing algorithm [6] has been inserted in the PSF general scheme of forecasting, allowing the new hybrid methodology to predict outlier occurrence [4]. Note that the goal of combining both techniques is not to provide accurate peak forecasts but to inform that the occurrence of an outlier is likely to happen. This methodology has been applied to 36 energy-related datasets, reaching a sensitivity of 86.97%, and a specificity of 80.46%, on average. The capability of detecting accurately outliers led to a forecast error decrease of almost 20%, since the forecast of these samples was avoided.

3. Conclusions

A new forecasting algorithm, based on clustering techniques, has been tested on several real-world time series showing its capability to provide accurate predictions for any kind of temporal data. In addition, the prediction of outlier occurrence has been successfully addressed.

Currently, different lines of work are being addressed. However, the discovery of motifs in discrete data to enhance the outliers prediction highlights among all. Equally remarkable is the way of computing the pattern sequences found in the historical data since density-based estimation are now being tested to improve the PSF accuracy. Last, how the use of evolutionary computation could improve PSF is also being analyzed.

Acknowledgement

The research was supported by the Spanish Research CICYT under project TIN2007-68084-C-02.

References