Robert Ulbricht, Hilko Donker, Claudio Hartmann, Martin Hahmann, Wolfgang Lehner

Challenges for Context-Driven Time Series Forecasting

Erstveröffentlichung in / First published in:

*Journal of Data and Information Quality. 2*016, 7(1-2), Art. Nr. 5 [Zugriff am: 07.10.2022]. ACM Digital Library. ISSN 1936-1955.

DOI: https://doi.org/10.1145/2896822

Final edited form was published in "Journal of Data and Information Quality" 17 (1-2), Art. Nr. 5. ISSN: 1936-1955
https://doi.org/10.1145/2896822

Challenges for Context-Driven Time Series Forecasting

ROBERT ULBRICHT and HILKO DONKER, Robotron Datenbank-Software
CLAUDIO HARTMANN, MARTIN HAHMANN, and WOLFGANG LEHNER,
Technische Universität Dresden

Predicting time series is a crucial task for organizations, since decisions are often based on uncertain information. Many forecasting models are designed from a generic statistical point of view. However, each real-world application requires domain-specific adaptations to obtain high-quality results. All such specifics are summarized by the term of context. In contrast to current approaches, we want to integrate context as the primary driver in the forecasting process. We introduce context-driven time series forecasting focusing on two exemplary domains: renewable energy and sparse sales data. In view of this, we discuss the challenge of context integration in the individual process steps.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—*Statistical databases*

General Terms: Time Series, Forecasting, Context

Additional Key Words and Phrases: Uncertain data, renewable energy, sales data, model selection, forecast evaluation

ACM Reference Format:
Robert Ulbricht, Hilko Donker, Claudio Hartmann, Martin Hahmann, and Wolfgang Lehner. 2016. Challenges for context-driven time series forecasting. J. Data and Information Quality 7, 1–2, Article 5 (April 2016), 4 pages.
DOI: http://dx.doi.org/10.1145/2896822

5

1. INTRODUCTION

In today's dynamic and competitive business environments, more and more decisions in planning and operations have to be made based on uncertain data. The ability of generating precise numerical forecasts becomes increasingly important for successful enterprises and is seen as one of the core tasks for modern data analysts. During the forecasting process, three aspects of data quality have to be considered: the quality of the available history for the target variable to be predicted, the quality of additional external influences (features), and finally the quality of the obtained output, thus denoting the forecast value itself. Applications can be found, for example, in the renewable energy domain or the retail industry: Forecasting the future energy supply from fluctuating sources like solar and wind power strongly depends on the quality of the

This work was supported by the European Regional Development Fund (EFRE) and the Free State of Saxony.
Authors' addresses: R. Ulbricht and H. Donker, Robotron Datenbank-Software GmbH, Stuttgarter Str. 29, 01189 Dresden, Germany; emails: {robert.ulbricht, hilko.donker}@robotron.de; C. Hartmann, M. Hahmann, and W. Lehner, Technische Universität Dresden, Dep. of Computer Science, Institute for System Architecture, Database Technology Group, 01062 Dresden, Germany; emails: {claudio.hartmann, martin.hahmann, wolfgang.lehner}@tu-dresden.de.

DOI: http://dx.doi.org/10.1145/2896822

Final edited form was published in "Journal of Data and Information Quality" 17 (1-2), Art. Nr. 5. ISSN: 1936-1955
https://doi.org/10.1145/2896822

available numerical weather model, as most of the output can be explained by a location's specific environmental conditions. This leads to a two-step approach, typically requiring an energy and a weather forecast model. Depending on the quality of input data, both of them will naturally include prediction errors. Having knowledge about the expected weather data quality might influence the selection of an appropriate statistical energy forecast model. Furthermore, the model can be refined by removing outliers resulting from measurement failures in the preprocessing and by using additional information like the installation's technical parameters in the postprocessing phase. In contrast to the energy domain, for sales data the influence of external factors on different product groups may not be expressed by numerical values. For example, big sports events like the Soccer World Cup or Olympic Games lead to an increased number of TVs sold, but it is not possible to transform that knowledge into a statistical model. A second problem is the extreme data sparseness of some product groups in the sales domain. Items that are sold well in one season do not sell well in the next season or even not at all and certainly not in all stores monitored. The most sparse datasets may have a coverage of only 1% of all possible items and store combinations over time. Usually, a meaningful integration of such external factors in the generic forecasting process requires access to expert knowledge and manual adjustments, thus increasing its complexity and slowing it down. On the other hand, neglecting such information would make the process too stringent and unable to adopt to specific domain requirements. For example, based on the generic reuse concept from Becker et al. [2007], context-aware parameter estimation is used to speed up the creation of energy demand models [Dannecker et al. 2011] while maintaining acceptable result quality.

2. CHALLENGES AND RESEARCH DIRECTIONS

By following the steps usually conducted in the forecasting process [Box et al. 2008] and considering the goal of context integration, these research challenges can be derived:

(1) *Handling Dynamic Context.* Context information can occur in different forms and formats; for example, explicitly modeled as rules or ontologies, implicitly extracted as correlations or stored in a case base. In order to process and store these various kinds of context information, a flexible integration process and a data model is required. The relevant parts of the incoming data must be identified and have to fit into appropriate data structures in order to make them usable. Due to its dynamic nature, context information must be assessed and monitored for changes constantly. With this certainly being a challenge for any relational data model, concepts like Schema Flexibility [Voigt and Lehner 2014] address such issues and should be given consideration during the application design.
(2) *Data Preparation.* Meaningful preparation of data requires two stages: quality enhancement and feature selection. The first stage increases raw data quality by handling issues like missing or incomplete input data [Strong et al. 1997]. As the point at which data is considered as low quality is context dependent, the integration of context/domain information into quality enhancement will increase its effectiveness. The second stage utilizes context information to identify and extract the most relevant features for the analysis technique to be applied and the domain from which the data originates. In diverging domains like energy production and retail sales, attribute names will most likely be orthogonal and even if they are not, they might describe different things. For instance, we know that for solar energy supply forecasts all night values can be ignored as they will always be zero, and the clear-sky model (maximal power output under perfect conditions) derived from historical observations is one important input feature, but none of this will work for

Final edited form was published in "Journal of Data and Information Quality" 17 (1-2), Art. Nr. 5. ISSN: 1936-1955
https://doi.org/10.1145/2896822

wind energy. Further, as different analysis algorithms can have different requirements for data quality, both phases of data preparation must be synchronized.

(3) *Model Selection / Adaption.* Identifying the optimal forecasting model for a specific problem can be an expensive task whenever there are different methods available. Currently, greedy trial-and-error approaches are used but by creating an extensive case base (e.g., using sample data from M3-competition [Makridakis and Hibon 2000] or other open data sources) and having context information mapped to the best known models, these efforts can be reduced and better results can be obtained in shorter time: Before a method is selected, the case base is queried automatically in order to find similar forecasting tasks and the models from the most suited cases are initially used as best practice and can be further refined. Hereby, the problem is the two-dimensional character of similarity: the similarity of the target time series itself, usually determined by distance measures and/or properties like standard deviation, signal-to-noise ratio, kurtosis and skew, and the likeness of associated context information; for example, a wind mill's geographic position or specific product properties like brand or color. Regarding data quality, specialized context- and quality-aware forecasting algorithms are needed to handle incomplete or low-quality datasets and in order to decide autonomously how much information is necessary to obtain optimal results. A first approach toward this direction has been proposed by Hartmann et al. [2015].

(4) *Output Evaluation.* Competing models are usually evaluated against naive reference models or using statistical accuracy criteria like the Root Mean Square Error (RMSE). Although easy to compute, such theoretical approaches might be seen equally hard to interpret and have limitations; for example, their inability of reflecting varying economic impacts of over- and underpredictions at a certain moment. By calculating a forecast's value based on real-time energy prices or the impact of over- and underproduction and storage costs for a certain retail product, a complex optimization problem is created leading to a far more realistic what-if analysis. We think that the definition of such a context-sensitive evaluation methodology, based on business rules and the derived monetary consequences, is a promising approach to avoid misleading model selection decisions.

Solving the problems described previously can open the door toward the development of a new class of expert systems, providing efficient support for an organization's daily forecasting tasks as well as the ability of conducting complex benchmarks to evaluate new forecasting methods. Achieving the goal of generic context integration can provide increased data quality, a more efficient forecasting process, and finally, better results.

REFERENCES

Jörg Becker, Christian Janiesch, and Daniel Pfeiffer. 2007. Towards more reuse in conceptual modeling—A combined approach using contexts reuse in conceptual modeling. In *Proceedings of the Conference on Advanced Information Systems Engineering Forum*. 81–84.

George E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. 2008. *Time Series Analysis: Forecasting and Control* (4th ed.). John Wiley & Sons Inc.

Lars Dannecker, Robert Schulze, Matthias Böhm, Wolfgang Lehner, and Gregor Hackenbroich. 2011. Context-aware parameter estimation for forecast models in the energy domain. *Lecture Notes in Computer Science*, Vol. 6809 (2011), 491–508.

Claudio Hartmann, Martin Hahmann, Frank Rosenthal, and Wolfgang Lehner. 2015. Exploiting big data in time series forecasting: A cross-sectional approach. In *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DASAA)*. 1–10.

Spyros Makridakis and Michele Hibon. 2000. The M3-competition: Results, conclusions and implications. *International Journal of Forecasting* 16 (2000), 451–476.

Final edited form was published in "Journal of Data and Information Quality" 17 (1-2), Art. Nr. 5. ISSN: 1936-1955
https://doi.org/10.1145/2896822

Diane M. Strong, Yang W. Lee, and Richard Y. Wang. 1997. Data quality in context. *Communications of the ACM* 40, 5 (1997), 103–110.

Hannes Voigt and Wolfgang Lehner. 2014. Flexible relational data model—A common ground for schema-flexible database systems. In *Proceedings of the 18th East European Conference on Advances in Databases and Information Systems (ADBIS)*. 25–38.