Adaptive Hierarchical Forecasting

C. Engels [1,2], W. Konen [3,2]

[1] *Fachhochschule Dortmund, Emil-Figge-Str. 42, 44227 Dortmund*

[2] *NuTech Solutions GmbH, Martin-Schmeisser-Weg 15, 44227Dortmund*

[3] *Fachhochschule Köln, Am Sandberg 1, 51643 Gummersbach*

***Abstract* - This paper describes the extension of classical forecasting methods for an application to hierarchical data structures. We show that various methods of hierarchical coupling can improve forecasting results using hierarchical relationships considerably. In a concrete application we are able to reduce the relative error of a retail forecasting model by 10%.**

I. INTRODUCTION

During the last decade Intelligent Data Analysis and Data Mining Methods have reached applicability in a variety of business domains. Predictive modeling has been integrated in many analytical CRM applications [1,2]. We find pattern detection and monitoring tools in internet security [3] and the identification of association rules in market basket analysis domains like Amazons wishlist to mention only some breakthroughs.

Standards have been established to guide the process starting from the specification of the business objectives, the estimation of data driven models and their application in the field of the business domain [4,5].

Beside these remarkable successes the central business question of demand forecasting has been neglected. The potential of high accuracy demand forecasting methods is getting more and more obvious since markets change their characteristics from supply driven to demand driven behavior. Optimization techniques [20] for product planning, production scheduling, inventory control and yield management rely on the estimation of future demands. If corporation success depends on the precision of demand forecasting what is the reason for being there not as mature as in the other Data Mining and Data Analysis process steps mentioned above? We assume that the complexity of the entity to be predicted namely the product structures have been the main obstacle for a more mature implementation of demand forecasting techniques. In practice we find complex hierarchical data dimensions which build up the product structure in most cases. In those sectors where product structures are more simple (e.g. as in the airline and hotel industry) compared to other industries we find successful applications of demand forecasting methods since twenty five years [6,7].

We provide a new methodological approach to expand the application of demand forecasting techniques beyond the level of individual and proprietary construction. This approach is called Adaptive Hierarchical Forecasting which combines already established and innovative methods in an appropriate process model. The resulting model extends the well known standards like the CRISP-DM or SEMMA model by adding the aspects of hierarchical data structures and adaptivity to the life cycle steps. Following this strategy we try to bring the application of demand forecasting from art to engineering.

II. APPROACH

Based on fifteen years professional experience within the field of demand forecasting we have identified the critical parameters which describe the complexity of product families, products and their variants, see Table I.

TABLE I.
PARAMETERS DESCRIBING COMPLEXITY

Parameter	Complexity
Lifetime of a product variant in terms of periods of existence	Fashion items are more complex than utility items
Volatility and seasonality of the demand	High volatility increases complexity
Depth of hierarchy structure (= number of hierarchical levels) for the description of a product variant	Deep structures lead to higher complexity
Number of observations (= transactions) per hierarchy level	Few observations are in conflict with the “Law of great numbers” in statistics
Business processes	Brochures, promotions, … stability of underlying business processes

Based on these parameters we are able to identify the hierarchy of products, see Fig. 1.

The classical approach is to identify one level within the hierarchy and to estimate a predictive model which generates a forecast for the total demand on this level. In contrast to this strategy we try to build several models on different levels due to the data characteristic on each level and try to build an integrated model among these models by mutual influence.

If we investigate forecasting models on different levels within a hierarchy we observe two major dependencies: dependencies between levels and relationships within levels (shift effects). In classical prediction modeling identification is based on a variety of driving inputs (e.g. input to neural networks) but those models are uncoupled. In contrast our approach is shown in Fig. 1: each circle depicts a forecasting model on a specific hierarchy level. In the top hierarchy levels models can be trained based on a large number of observations which allow higher forecasting precision but lower granularity. Fig. 1 shows two examples which have been highlighted: some circles (black) depict forecasting models which have limited prediction capabilities due a small number of observation patterns. The coupled process model allows for

stronger influence from upper level models (top-down). The grey circle shows a model which operates in the inverse direction in a consistent way (bottom-up). This model is driven

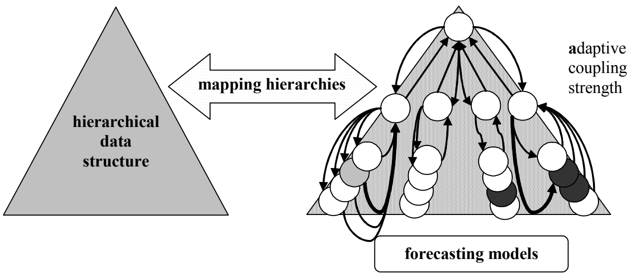

Figure 1: Forecasting as a coupled process

by explicit knowledge about the business processes, e.g. promotional actions for a dedicated product variant.

Partially this approach has been implemented in different real world applications in the tour operator, energy and distant selling sector [8,9]. Based on incremental neural networks for convenient setting of model parameters [10,11] these models calculate several thousand demand values each day efficiently [9,12] and adapt in real time to one shot events. The applied techniques enable the prediction in total values a well as in relative portions of an aggregated value one aggregation level above. Therefore the resulting model can be simulated and analyzed [13,14].

III. Architecture

We are able to apply four different approaches to integrate the different levels within a hierarchical architecture: First we distinguish explicit and implicit coupling:

A. Explicit Coupling

Explicit coupling generates independent predictions on each level and merges these results via explicit formulation. In the **top-down direction** the upper level output $\hat{y}_p$ is distributed to the lower level. In Eq. (1) the predicted distribution factors $\hat{y}_v^0$ of a product variant on the lower level are normalized to a total of 1, see $\hat{y}_v^{norm}$. With Eq (2) we constrain the sum of product variants $\hat{y}_v$ to the total product count $\hat{y}_p$.

$$\hat{y}_v^{norm} = \hat{y}_v^0 / \sum_{V_P} \hat{y}_v^0 \quad (1)$$

$$\hat{y}_v = \hat{y}_p * \hat{y}_v^{norm}, v \in V_P, p \in P \quad (2)$$

Alternatively, we can use the **bottom-up direction** to generate the product prediction, or to find a model bias, see Eq. (3). This bias can be used to correct the prediction of upper level models.

$$\hat{y}_p = \sum_{V_P} \hat{y}_v \quad (3)$$

$\hat{y}_v$ denotes the derived prediction value of the stock keeping units (SKU) of product variants while $\hat{y}_v^{norm}$ gives the prediction of the relative portion of a product variant in relation to the product in total in percent.

B. Implicit Coupling

In principle there are two ways to build the lower level model: a) it receives input from the same aggregation level which we call **intralevel input** (see Tab.II (A)) or b) we add higher aggregated input variables from the levels above which we call **interlevel input** (see Tab.II (C)).

In the second case we will have to find a balance between increasing the number of input dimensions and prediction results since not all estimators are able to cope with high dimensional input.

IV. Experimental Results

The task we have chosen in this paper is forecasting real demand data from the retail sector on the upper article (in the following called product) level P and a lower level which contains variants of the product level V. For sake of simplicity the problem has been reduced to two levels only without loss of generality. The set of product variants which belong to one product is denoted as V_p.

Products are ordered on a semiannual seasonal basis and the total demand for one season has to be forecasted. The product portfolio is changed entirely from season to season.

36% of the raw product level (P) data records contain missing values whereas this number jumps to 85% on the variant level V. Most of the missing values are observed within the leading indicator input variables.

We have compared recent variants of neural networks [17] and random forests [19] on these data to evaluate adaptive hierarchical forecasting architectures. This paper concentrates on the random forest results because random forests exhibit well reproducible results in runs with different random seeds (+/- 1% in residual value) while the neural nets varied over a residual range of +/- 3%.

We also found the random forest characteristics already described by Breiman [19] to be valid on the given datasets:

- excellent accuracy
- scales up
- handles thousands of variables
- many valued categoricals
- extensive missing values
- badly unbalanced data sets
- gives internal unbiased estimate of test set error
- cannot overfit
- variable importance
- outlier detection
- data views via scaling

TABLE II.
LOWER LEVEL PREDICTION ERRORS (MAPE) FOR VARIOUS MODEL COUPLINGS

	Lower level model			
Upper level model	**(A) Intralevel input only**	**(B) Interlevel and intralevel input (excluding upper level output)**	**(C) Explicit coupling (including upper level output)**	**(D) Implicit coupling (including upper level output)**
No model	**62,77%**	**60,78%**		
Intralevel input			**66,35%**	**56,22%**

It turns out (not shown here) that the explicit approach is promising if the residuals on the upper level are much smaller compared to the errors on the lower level. Tab. II shows MAPE errors on the lower level on the retail demand data set using the explicit approach according to (1) and (2) with (C) and without (A) interlevel constraining.

If the upper level residuals are not smaller than the residuals on the lower level by an order of magnitude we prefer an implicit approach where the lower level model also receives the selected input and output from the upper level model. In the example retail data of Tab. II the prediction error on the upper level was 44%. In this case the implicit approach should be better which is validated by Tab. II: It shows results with (D) and without (B) interlevel output. By this choice of an appropriate coupling mechanism we were able to reduce the prediction error by 62.77% - 56.22% = 6,5% (Tab.II, first row in (A) vs. last row in (D)).

The implicit approach needs more input dimensions in the lower level model and we have to use estimators like [19] where high dimensional input doesn't lead to an increase of prediction errors. We use 22 input variables on the upper and 17 input variables on the lower level where in the latter case 4 input variables represent aggregated input values from the upper level.

V. CONCLUSION

We have shown that a coupling between hierarchical levels leads to increased prediction performance. However the mode of coupling turns out to be critical: While explicit coupling in certain circumstances can deteriorate model performance slightly, the implicit coupling mode can boost the model performance significantly (10% decrease in relative error)

In a next step we are currently automating and integrating these approaches to bring demand forecasting dealing with hierarchical data structures one step ahead. This work includes the interplay of multiple aggregation levels. There we find the combination of the explicit and implicit approach since top level models usually exhibit small prediction residuals due to the law of great numbers.

Innovative model estimators are under further investigation [15,16,17,18,19] and we plan to give a detailed description of the selection and adaptation processes in terms of an extended data mining life cycle process model.

ACKNOWLEDGMENT

We thank Sebastian Lux for his major contributions to the hierarchical approach and so many fruitful discussions. We also thank Gunnar Lotz for data preparation and modeling and Markus Skotarzik for his help with the R-coding and random forests. Further we want to express our gratefulness to Thomas Bäck and Ulrich Hammel for building such a warm and constructive atmosphere at NuTech.

REFERENCES

[1] http://www.nutechsolutions.com/petro_mkt.asp
[2] http://www.sas.com/offices/europe/germany/download/files/solutions/SAS_Enterprise Miner_Broschuere.pdf
[3] J. Seifert, Data Mining and Homeland Security: An Overview, http://www.fas.org/sgp/crs/intel/RL31798.pdf
[4] http://www.crisp-dm.org/
[5] http://www.sas.com/technologies/analytics/datamining/miner/semma.html
[6] http://www.manugistics.com/solutions/
[7] http://www.prosrm.com/new_content/Industry_Solutions.htm
[8] C.Engels, Prognose für Versorgungsunternehmen, Whitepaper, Thinking Networks SI, Bochum, Februar 2001.
[9] U. Mohr and C. Engels, Yield Management für Flugpauschalreiseveranstalter 2. Conference der Neuro Fuzzy Initiative, Düsseldorf, 1997.
[10] C. Engels: Incremental Neural Gas. Proceedings EUFIT `95, Aachen, 1995.
[11] C. Engels: Vorhersage von Wirtschaftsdaten mit neuronaler Technologie. 3. Symposium Softcomputing, Göttingen, 1997.
[12] C. Engels: Revenue Management beim weltgrößten Reiseveranstalter: Der Kollege "Prognose". IS Report, Zeitschrift für betriebswirtschaftliche Informationssysteme, Vol. 10, Okt. 2001.
[13] W. Konen: Ertragssteigerung durch integrierte Prognose und Planung, SAS-DISK (20. deutschsprachige SAS-Konferenz), Mannheim, Nov. 2002.
[14] W. Konen: Optimieren durch Simulieren: Neue Steuerungslogik wird mit Echtdaten getestet. IS Report, Zeitschrift für betriebswirtschaftliche Informationssysteme, Vol. 10, Okt. 2002.
[15] R. Caruana, Multitask Learning, Machine Learning, 28, pp. 41-75, 1997.

[16] R. Caruana, A Dozen Tricks with Multitask Learning, in: G.B. Orr, K.-R. Müller (eds.) Neural Networks: Tricks of the Trade, 1998.
[17] M. Feindt, A Neural Bayesian Estimator for Conditional Probability Densities, http://arxiv.org/pdf/physics/0402093.
[18] E. Reinus, Vergleichende Untersuchung von Multi-Task-Learning zu herkömmlichen neuronalen Netzen am Beispiel der Prognose, Diploma Thesis, University of Applied Sciences Cologne, 2005
[19] L. Breiman, Random Forests, Technical Report for Version 3, University of California, Berkeley, CA, 2001.
[20] T. Bäck: Evolutionary Algorithms in Theory and Practice. Oxford University Press, New York, 1996.