2009 IEEE International Conference on Data Mining Workshops

Towards Context Aware Food Sales Prediction

Indrė Žliobaitė, Jorn Bakker, Mykola Pechenizkiy
Eindhoven University of Technology
P.O. Box 513, NL-5600 MB, Eindhoven, the Netherlands
zliobaite@gmail.com, {j.bakker, m.pechenizkiy}@tue.nl

Abstract

Sales prediction is a complex task because of a large number of factors affecting the demand. We present a context aware sales prediction approach, which selects the base predictor depending on the structural properties of the historical sales. In the experimental part we show that there exist product subsets on which, using this strategy, it is possible to outperform naive methods. We also show the dependencies between product categorization accuracies and sales prediction accuracies. A case study of a food wholesaler indicates that moving average prediction can be outperformed by intelligent methods, if proper categorization is in place, which appears to be a difficult task.

1 Introduction

Demand prediction is essential part of business planning. An accurate and timely sales prediction is very important for stock management and profitability. In food sales the stock includes large assortment of goods, some of them require special storage conditions, some are quickly perishable.

Challenges of sales prediction. There are general and product specific causes of the demand fluctuation. The variations in consumer demand may be influenced by a price change, promotions, changing (rapid or gradual, global or local) consumer preferences, or weather changes [6]. Furthermore, seasonal changes occur due to different cultural habits, national and religious holidays, or fasting. All these factors imply that some types of products have high sales during a limited period of time.

Although seasonal patterns are expected, the predictive features that define these seasons are not always directly observed. Besides, the historical data is often highly imbalanced with only a few peaks per year.

Variation in sales figures can be classified into short term fluctuations (e.g. party today, shopping tomorrow), medium term seasonal patterns (e.g. June vacations) and long term trends (e.g. economic situation). Different horizons of the *food* sales predictions are required for business decisions. We focus on medium term seasonal patterns (weekly predictions), which are essential for stock management.

Related work and our approach. Moving averages of different lag or simple regression models are often used basic approaches. In this setting baseline predictions are often overridden by managers using their intuition and expertise. Predictions based on moving averages may work well when demand is flat. But when the demand follows trends or seasonal patterns the reaction of moving average is too slow. Managers often try to improve the performance in seasonal peak periods by prudently increasing the stock and thus costs.

Another typical approach is to have a number of reminders that should hint about the coming holidays, weather and other demand triggers. However, human factors like overload of information, lack of expertise (new personnel), or forgetfulness may result in mistaken predictions and poor decision making.

We present a context aware prediction approach (CAPA) for food wholesales[1]. First, we learn how to categorize the sales time series offline into four categories ("flat", "frequent", "occasional" and "seasonal") based on their structural properties. Next, for each product classified to a particular category, we online apply a prespecified base predictor.

Our approach comes close to meta learning [7], where the relevant learners are selected based on offline predefined criterion. In this study we introduce a generic sales prediction approach with context awareness. In contrast with the meta learning approaches [8, 5], we incorporate domain expertise and observations in categorization and base predictor selection process.

One could argue, that an ensemble approach does that automatically. All possible input features can be collected

[1] An extended version can be found as a Technical Report `http://www.win.tue.nl/~mpechen/projects/sligro/`

978-0-7695-3902-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDMW.2009.60

and then apply rigorous feature selection and predictor selection from an ensemble. This approach has limitations with respect to a food sales prediction problem due to relatively short available sales history and common patterns (e.g. New Year). By categorizing the time series based on their structural properties, we expect to narrow down the job for the particular predictor, allowing to focus on the peculiarities of a particular series.

Another related approach would be a multi task learning approach [1]. Yet instead of taking a "black box" approach we use explainable judgment to filter out relevant part of the task.

In the following sections we consider the most essential details of our CAPA approach, and then present a case study of Sligro Food Group N.V., the food wholesaler.

2 Context aware sales prediction approach

The main idea of CAPA is to select the predictor based on the structural properties of historical time series. If we can identify and extract distinct categories of products, specific input data construction procedures and specific predictors could be employed for each category.

In this section we describe both the offline and online operation of CAPA. The offline part of CAPA consists of extracting structural features from the time series and categorizing the time series according to different types of behavior. The online part based on this categorization selects the best predictor model for a given product. We do not focus on the actual prediction procedure here.

2.1 Decision support with CAPA

CAPA consists of two blocks - training (offline) and operating (online). First let us assume that the model has already been trained offline, i.e. the categories have been fixed, mapping of time series to the categories is established, and "local" expertise of each predictor is known.

Figure 1 presents online operation of CAPA. Let us take a particular product we are interested to predict online. First of all we extract structural features from the original sales time series of the product. Then we assign the product to one of the categories. We pick a base predictor specific to a particular category and select input features, relevant to a particular category. That is the context aware part of the approach. The contexts are specified by predefined categories. Having the original series, the base learner, and the input features we can cast the prediction.

The CAPA approach is generic and we could emphasize here a distinction between context in time and context in space. However, for simplicity we assume that a particular product does not migrate from one category to the other, thus leaving context in time outside the scope of this study.

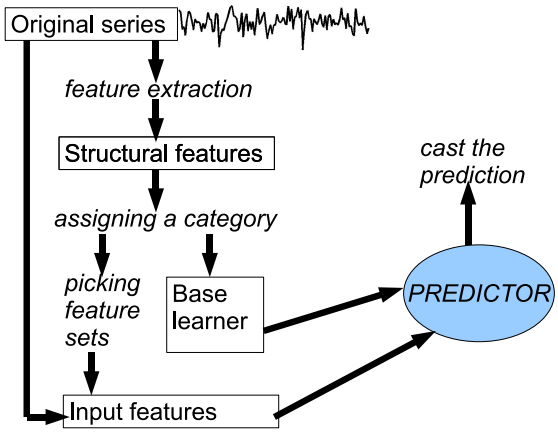

Figure 1. Online operation of CAPA.

2.2 Training CAPA

The core part of context aware prediction approach is to match the product categories and the base predictors.

Offline part. A limited set of base predictors $(G_1, G_2, \ldots, G_m)$ needs to be preselected based on domain knowledge and expectations in order to delimit the full state space search. Then for each product m parallel experiments are carried out using each of the predictors. Next, the products are grouped into m categories based on the best performing predictor. Each obtained category serves as a basis for constructing categorization rules.

Online part. The goal of the training process is to learn to assign a product to one of these defined categories online, having only a fragment of the series. When we have the categorization rules, an unseen product can be processed as described in Figure 1. First of all the category of the product is determined, say C_j. Then the corresponding predictor G_j is used to output the prediction.

2.3 Structural features

The aim is to identify different types of sales behavior and use this information in picking a predictor. In order to capture behavior in time, we consider different structural properties of the time series that reflect the behavior in time. This set of structural features is then the context from which new time series are identified.

Using external knowledge and visual observations, we can categorize the series using two dimensions: seasonality and deviations (see Figure 2 with examples of artificial series, for illustration purposes).

For example, bread sales can behave like "flat" series. "Frequent" series represent the products, which are bought in large quantities, following no particular seasonality. "Occasional" product sales increase sharply in relation to par-

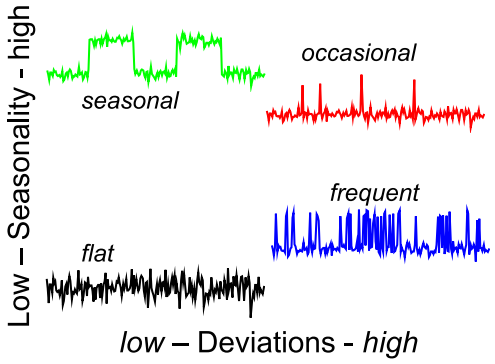

Figure 2. Considered types of behaviors

ticular occasions, like eggs for Easter. Ice cream is a "seasonal" products with respect to weather.

We normalize the values of the series to be in a range $(0, 1)$ before defining structural features. We extract the following structural features:

- F_{s1-s3} |mean - median|, standard deviation, shift;
- F_{s4-s8} threshold h crossing ratios;
- F_{s9-s10} normalized power of the frequency p in the frequency spectrum;
- $F_{s11-s12}$ local variation features: interquartile range and unequal neighbors.

The features F_{s1-s3} capture global characteristics of the series y. Shift is the mean number of points for which $y_t < \mu$ minus the median of the number of points for which $y_t < \mu$, where μ is the mean of the time series.

In order to capture the structural behavior of time series, we define a number of threshold values and note the number of times the signal crosses these thresholds (features F_{s4-s8}). This is done for the threshold values $h = 0.5 \pm 0.1, 0.7, 0.3, 0.8, 0.2$. This number is then scaled to the total length of the signal to obtain the ratio.

Seasonal patterns could manifest themselves as, for instance, yearly or bi-yearly changes. This information should appear in the frequency spectrum of the time series as a relative high power in the frequencies $p = 1/52$ and $2/52$. In order to obtain these features we apply Fast Fourier Transformation (Cooley-Tukey implementation [2]) and extract the corresponding frequencies (features F_{s9-s10}).

We aim to capture local variation using features $F_{s11-s12}$. Unequal neighbors is the mean number of times $y_t \neq y_{t-1}$. The interquartile range of $y_t - y_{t-1}$ is a robust measure for the spread of variation in the signal. Any outliers that fall in the upper or lower 25% of the difference distribution will not affect it.

2.4 Learning categorization rules

We would like to learn categorization rules, which would assign a given product to one of the defined categories based on it's structural features.

We take two approaches, which we call *bottom up* and *top down*. In the first approach we use the training accuracies to label the training products and using these labels try to learn categorization in a supervised manner. In the second case we visually pick a set of representations from the four categories ("flat", "frequent", "occasional" and "seasonal") defined earlier and use them as prototypes to learn the categorization.

2.4.1 Bottom up categorization

In order to test whether the category assignments are learnable, we train a classifier on the "true" labels of the categories generated using the training dataset. The labels are obtained by running all the classifiers from the pool for all the products and then ranking the accuracies for each product. A product gets the label, corresponding to the best performing classifier.

If the categorization is learnable we should be able to assign an optimal predictor based on the structural features.

Training. A decision tree classifier is used to make the mapping from structural features to classes. We use the 12 structural features described in Section 2.3. For every possible subset of features a classifier is trained and evaluated. For each of the classifiers the mean accuracy of the intelligent predictors is used as an evaluation measure.

Validation. The best classifiers should be able to generalize the classification. We use cross validation on the training set to select the categorization rule.

2.4.2 Top down categorization

We visually pick a number of products to represent a category. Then we extract structural features (defined in Section 2.3) from each of the picked series. We average the structural features within each of the four categories and the averages serve as the four prototypes. Finally, we cluster the products to the four categories, using prototypes as fixed cluster centers.

2.5 Input space features

We already discussed the extraction of structural features, which are used for time series categorization (see Figure 1), now we present the input features.

The input feature space used by the prediction models is formed using internal and external data. The internal data comes from a company sales database (Figure 3). The external data (holidays, temperature, seasons) is formed using

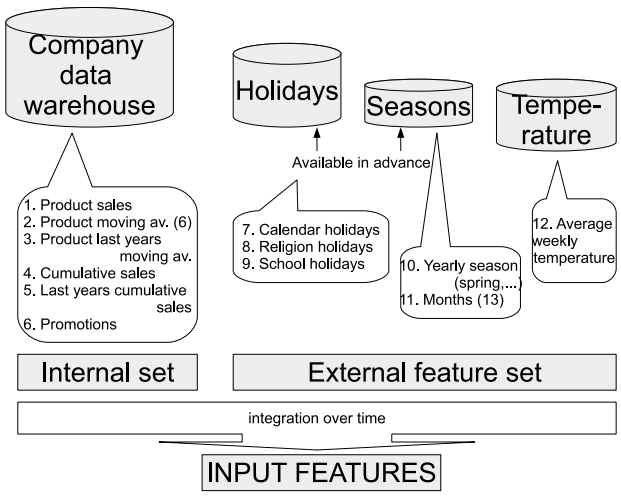

Figure 3. Formation of the input space for learning predictors

information from the ministry of culture, meteorological institute and common knowledge.

The internal features are interrelated. Moving average (F_{p2}) is calculated using (F_{p1}). Last years moving average (F_{p3}) is formed from (F_{p1}) using one year lag. Cumulative sales (F_{p4} and F_{p5}) include the sales quantity of all the products. Promotions (F_{p6}) for some products are organized.

The external features (F_{p7-p11}) are available in advance. Holidays (F_{p7-p9}) are described in 16 binary features. Seasons (F_{p10}) are described in 4 binary features, and months (F_{p11}) are described in 13 binary features. Temperature (F_{p12}) is described in a single numerical dimension.

3 SLIGRO case study

In this section we study a case of Sligro Food Group N.V. (SLIGRO) sales prediction. The company is engaged into food wholesales. SLIGRO works with corporate clients, mainly food retail and food service companies (restaurants), although there are some direct consumers as well. SLIGRO has around 40 outlets in the Netherlands. SLIGRO trades about 60000 products.

3.1 Prerequisites

We aim to test if the base predictor, which is selected based on time series category, consistently outperforms alternative models in terms of prediction accuracy.

Our experimental field consists of 538 product sales quantities over two years period (from July 2006 to October 2008). The products were selected based on sales volume. The sales are aggregated on weekly basis, thus each series is of 119 weeks length. Each series represent the sales of one product aggregated over all outlets.

We use a regression as the base classifier, equipped with feature selection mechanism. We run Principal Component Analysis for feature reduction and keep the new features, which explain at least 70% of the data variance.

We use discretized labels to achieve comparability between different products. We discretize the outputs to 8 classes from very low sales (1) to very high (8), using Symbolic Aggregate Approximation (SAX) [4].

For model evaluation we use *Mean Absolute Scaled Error* (MASE) [3]:

$$MASE = \frac{1}{n}\Sigma_{t=1}^{n}|\frac{e_t}{MAE(Baseline)}|, \qquad (1)$$

where e_t is the prediction error at time t, $MAE(Baseline)$ is the mean absolute error of the baseline method. We use naive one step ahead prediction as the baseline (the prediction for the next week is equal to the factual sales this week). This is a moving average, when the lag is equal to one.

Using MASE as the accuracy measurement, we compare the intelligent classifiers to the baseline method.

3.2 Experimental set up

The experimental scenario consists of three parts: selection the base predictors, learning categorization rules, and testing final model accuracies. In the first part we select the base predictors and obtain "true" labels. We show that there exist product subsets on which it is *possible* to outperform baseline predictor. In the second part we aim to learn the dependencies between product categorization accuracies and sales prediction accuracies applying two approaches *bottom up* and *top down*. In the third part we test the final model accuracies.

To quantify the results, we perform controlled experiments using 538 product sales history. We take out random 100 series from the dataset and reserve them for final *model testing*. We call this set **T**. We develop the model using the remaining 438 series. We call the remaining set **M**.

3.2.1 Selection of the base predictors

We narrow down selection of the base predictors to a regression with different sets of input features, based on the observed categories (Figure 2). The input features for each predictor (MA(lag) is a moving average, reg. is regression) are listed in Table 1.

We run all the predictors on 438 series. We split the series into two parts 57 weeks used as a "warm up" and then the remaining 61 weeks are used for sequential testing. We obtain 438×7 matrix of scaled accuracies (MASEs). Then

Table 1. Base predictor selection

	Base	Input features
G_1	MA(1)	$F_{p1}^{(t)}$
G_2	MA(3)	$F_{p1}^{(t-5...t)}$
G_3	MA(6)	$F_{p1}^{(t-5...t)}$
G_4	reg.	$F_{p1}^{(t-5...t)}$, $F_{p6}^{(t+1)}$
G_5	reg.	$F_{p1}^{(t-5...t)}$, $F_{p2..p5}^{(t)}$,$F_{p6,p11}^{(t+1)}$,
G_6	reg.	$F_{p1}^{(t-5...t)}$, $F_{p2}^{(t)}$, $F_{p6,p10}^{(t+1)}$, $F_{p7..p9,p12}^{(t,t+1,t+2)}$
G_7	reg.	$F_{p1}^{(t-5...t)}$, $F_{p2..p5}^{(t)}$, $F_{p6,p11,p12}^{(t+1)}$, $F_{p9}^{(t,t+1,t+2)}$

we group the products based on the top ranking predictor. This way we get the "true categories" for given settings. In Table 2 the average MASEs for each of the "true categories" are provided.

Table 2. MASEs for the "true categories".

	Size	G_1	G_2	G_3	G_4	G_5	G_6	G_7
C_1	200	**1.00**	1.31	1.63	1.63	1.89	1.91	1.93
C_2	68	1.00	**0.89**	1.02	1.25	1.56	1.53	1.62
C_3	36	1.00	0.92	**0.85**	1.04	1.21	1.21	1.22
C_4	35	1.00	1.04	1.09	**0.85**	1.00	1.01	1.03
C_5	19	1.00	1.06	1.17	0.96	**0.86**	0.99	0.93
C_6	43	1.00	1.02	1.11	0.97	0.98	**0.85**	0.94
C_7	37	1.05	1.08	1.18	1.02	0.98	0.98	**0.90**

The best results appear on the diagonal in bold. The results below 1 mean that *it is possible to outperform moving average* if we do the correct categorization online. Furthermore, there are more than a single case per line, which outperform the baseline predictor.

3.2.2 Learning the categorization

We test the two categorization approaches *bottom up* and *top down* (Section 2.4). We delimit the task to four categories, which we expect to correspond to the product categories presented in Figure 2. Thus we group the three moving averages (G_1, G_2, G_3) into one category. A basic regression (G_4) forms the second category C_2. The third category C_3 is the calendar based regression (G_6). Regressions including annual patterns (G_5 and G_7) form the fourth.

In both cases we use the full length of the series (119 weeks), normalized to fit the range $(0, 1)$.

We present the results of the categorization procedure in the next section together with the final accuracies.

3.2.3 Prediction accuracies

We develop and validate CAPA model using the training set **M** and test assuming online settings on **T**. We split each series into "warm up" (57 weeks) and testing (61 weeks). For the dataset **M** we use the categorization, which was obtained during the training phase. We categorize the products from the dataset **T** using only "warm up" part and the categorization rules obtained in the training phase.

We run sequential testing of the four models G_1, G_4, G_6 and G_7, which we assume to correspond to each of the four product categories. We run the four models on each of the series from both datasets **M** and **T**. We compare the four accuracies for each product series. We aim to minimize the MASE for each pair of model-category.

First, we present the accuracies of the training data **M**, which corresponds to offline settings. In Table 3 average MASEs for the obtained categories are listed. It can be seen that in training *bottom up* approach (a) the selected base predictors G_4, G_6 and G_7 outperform the baseline predictor in the corresponding categories C_2**,** C_3 and C_4 .

To validate the results, we assign each product to a random category and then calculate average MASEs correspondingly. We present the results in the same Table 3 (c). In random categorization case the baseline predictor prevails. The *top down* categorization method (b) does not outperform the baseline predictor, however it gives better than random results, leaving prototyping approach as promising future direction. The poor performance of the method can be attributed to categorization accuracy, which is only 43% on the training data. However, random categorization would be only 25%. Thus we are better than random, but not enough to beat the final accuracies of the baseline predictor.

In Figure 4 we depict MASE of the four categories as a function of categorization accuracy. 100% corresponds to the "true categories", 0% to random categorization. In the area within the ellipse in the Figure 4 the three categories (C_2, C_3, C_4) outperform the baseline predictor. This shows the benefit of applying CAPA if we achieve at least 85% categorization accuracy.

Table 3 presents average MASEs of the obtained categories on the dataset **T** for *bottom up* (d) and *top down* (e) categorization approaches. Along we present the results of random categorization (f). The results do not show MASE below 1 for the target categories. This is due to not sufficient categorization accuracy, which is 47% for (a) and 43% for (b). However, random categorization would give only 25% accuracy. Thus we managed to learn some categorization and these are promising results.

Table 3. MASEs for (a,d) trained categorization *bottom up*, (b,e) trained categorization *top down*, (c,e) random.

		Size	G_1	G_4	G_6	G_7
Training accuracies (M)						
(a)	C_1	328	**1.01**	1.43	1.68	1.71
	C_2	34	1.00	**0.89**	0.97	1.00
	C_3	31	1.00	1.00	**0.94**	0.97
	C_4	45	1.00	0.98	0.95	**0.91**
(b)	C_1	158	**1.01**	1.69	2.13	2.20
	C_2	32	1.00	**1.34**	1.40	1.40
	C_3	7	1.00	0.95	**0.94**	0.92
	C_4	241	1.00	1.07	1.12	**1.12**
(c)	C_1	100	**1.01**	1.36	1.59	1.54
	C_2	108	1.00	**1.35**	1.66	1.59
	C_3	117	1.01	1.28	**1.60**	1.51
	C_4	113	1.00	1.35	1.63	**1.56**
Testing accuracies (T)						
(d)	C_1	69	**1.00**	1.37	1.57	1.65
	C_2	3	1.00	**1.41**	1.63	1.60
	C_3	20	1.00	1.62	**1.84**	1.85
	C_4	8	1.00	1.17	1.26	**1.28**
(e)	C_1	39	**1.00**	1.29	1.37	1.43
	C_2	4	1.00	**1.22**	1.39	1.42
	C_3	1	1.00	1.21	**1.23**	1.36
	C_4	56	1.02	1.32	1.67	**1.70**
(f)	C_1	24	**1.00**	1.30	1.50	1.59
	C_2	33	1.00	**1.37**	1.57	1.63
	C_3	15	1.00	1.72	**1.92**	1.93
	C_4	28	1.00	1.39	1.58	**1.63**

4 Conclusion

We developed CAPA, context aware sales prediction approach, via introducing background knowledge and visual observations into predictor selection process.

In SLIGRO case study we showed that distinct categories exist, where the intelligent learners can outperform naive predictors if online categorization is accurate enough (in SLIGRO case 85%).

We showed that is is possible to learn the "true", we obtained 47% accuracy on the testing set, while random categorization gives only 25% accuracy. However, we did not reach 85% accuracy which is necessary to have an advantage of intelligent methods at the final prediction. One of the reasons is that for online categorization only one year sales history is available, which does not generally allow for reoccurring contexts to appear.

Further improvement of the categorization accuracy could be achieved by finding more representative structural features, introducing multiple category assignments, adding more domain knowledge to the base predictor selection.

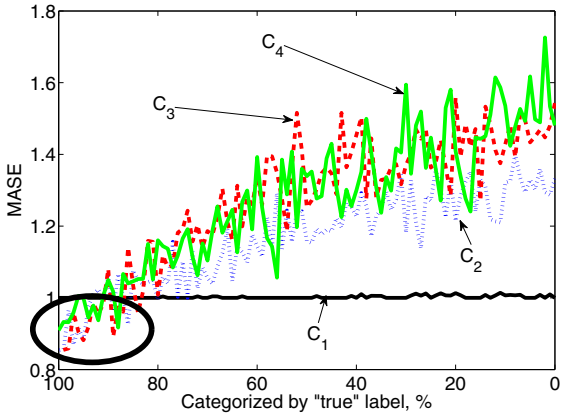

Figure 4. MASE and categorization accuracy.

Acknowledgements. This research is partly supported by NWO HaCDAIS project and LOIS grant. We are thankful to Sligro Food Group NV for providing us with the data and domain knowledge.

References

[1] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[2] J. W. Cooley and J. W. Tukey. An algorithm for the machine computation of the complex fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[3] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *Int. J. of Forecast.*, 22(4):679–688, 2006.

[4] J.Lin, E. J. Keogh, L. Wei, and S. Lonardi. Experiencing sax: a novel symbolic representation of time series. *Data Min. Knowl. Discov.*, 15(2):107–144, 2007.

[5] R. Klinkenberg. Meta-learning, model selection and example selection in machine learning domains with concept drift. In *Ann. Workshop on Machine Learning, Knowledge Discovery, and Data Mining (FGML-2005) Learning - Knowledge Discovery - Adaptivity (LWA-2005)*, pages 164–171, 2005.

[6] J. van der Vorst, A. Beulens, W. de Wit, and P. van Beek. Supply chain management in food chains: improving performance by reducing uncertainty. *Int. Transactions in Operational Research*, 5(6):487–499, 1998.

[7] R. Vilalta and Y. Drissi. A perspective view and survey of meta-learning. *Artificial Intell. Review*, 18:77–95, 2002.

[8] G. Widmer. Tracking context changes through meta-learning. *Machine Learning*, 27(3):259–286, 1997.