# Food Sales Prediction: "If Only It Knew What We Know"

Patrick Meulstee[a,b]
[a]Sligro Food Group
Corridor 11
PO Box 47 5460 AA Veghel
the Netherlands
pmeulstee@sligro.nl

Mykola Pechenizkiy[b]
[b]Eindhoven University of Technology
Department of Computer Science
P.O. Box 513, 5600 MB Eindhoven,
the Netherlands
m.pechenizkiy@tue.nl

## Abstract

*Sales prediction is an important problem for different companies involved in manufacturing, logistics, marketing, wholesaling and retailing. Food companies are more concerned with sales prediction of products having a short shelf-life and seasonal changes in demand. The demand may depend on many hidden contexts, not given explicitly in the form of predictive features. Even if some changes are known to be seasonal, predicting (and even detecting) when season will start and end remains to be non-trivial. In this paper we present an ensemble learning approach that employs dynamic integration of classifier for better handling of seasonal changes and fluctuations in consumer demands. We focus our research on studying how the business is currently operated, and how we can improve predictions for each product by constructing new groups of predictive features from (1) publicly available data about the weather and holidays, and (2) data from related products. We evaluate our approach on the real data collected by food wholesaling and retailing company. The results demonstrate that (1) our ensemble learning approach can perform better than the currently used baseline, (2) we can handle seasonal changes with ensemble learning better if feature set for a target product is complemented with features of related product (having similar sales pattern), and (3) an ensemble can become more accurate if information about the weather and holidays is presented explicitly in a feature set.*

## 1. Introduction

The success of different companies depends today on their ability to adapt quickly to the changes of their business environment. An accurate and timely sale prediction is in particular very important for the companies involved in manufacturing (improving planning and resource man-

agement), logistics (improving infrastructure and reducing inventories), marketing (assessing an impact of changes in a marketing strategy on sales), wholesaling and retailing (increasing profits and the service level).

In the food and beverages market, food service companies have to deal often with short shelf-life products, and uncertainty and fluctuations in consumer demands. These variations (whether seasonal or not) in consumer demand may be impacted by price (change), promotions, changing (rapid or gradual, global or local) consumer preferences or weather changes [19]. Furthermore, a large share of the products sold in that market is sensitive to some form of seasonal change due to the different cultural habits, religious holidays, fasting and alike. All these factors imply that some types of products are sold mostly during the limited period(s) of time.

Although it is known that some seasonal pattern is expected, the predictive features that define these season are not always directly observed. Therefore, drops and rises in sales which are accommodated by the changing seasons are often difficult to predict. Regarding inventory management, this results often in a stock-out at the start of the season and perishable or obsolete goods at the end of a seasonal period. Thus, both shortage and surplus of goods can lead to loss of income for the company.

A recent initiative of a few large companies in the food industry, which aimed to improve sales prediction practice, identified that 48% of food companies are not doing well in forecasting [1].

Time-series research has been traditionally suggesting ARIMA (autoregressive moving average) and ANN (artificial neural networks) approaches to address the problem of sales prediction. Despite of the continuous efforts devoted to come up with a right algorithm, and a number of comparative studies focused on identifying the strongest one, researchers are not clearly in favor of one particular method. Nonlinearity prevents the success of simple linear models, while rather short lengths of the time series are insufficient

to learn more complex models [3].

Several success studies have been reported in the literature. However, some surveys initiated by the several large companies in the food industry indicate that the majority of food companies still suffer from poor sales predictions, which lead to losses for style goods and perishable items, lost opportunities, and decrease of the service level [1].

In this paper we present the ensemble learning approach for sales prediction. This approach employs dynamic integration of classifier for better handling of seasonal changes and fluctuations in consumer demands. In recent studies it has been shown that dynamic integration outperforms other ensemble approaches on data known to have concept drift (that is the changes in the hidden context inducing more or less radical changes in the target concept [23]) and that it can be successfully used to predict antibiotic resistance [17].

We focus our research on studying how the food wholesale prediction is currently operated in a particular company, and how we can improve predictions for each product with our ensemble learning approach.

While we learnt that domain experts often exclaim "If only we knew what we know" after making yet another mistake in decision making processes, we suggested to change this at least to "If only it (i.e. decision support system) knew what we know (as domain knowledge)" with a hope that at the end it will result in the changes towards "Oh, it knew even what we still do not know at the moment".

Besides the analysis of what expert knowledge domain experts use for sales prediction and introducing this knowledge implicitly in the form of predictive features (from the publicly available data about the weather and national holidays), we introduce a generic idea of using external features from related objects. In food sales prediction case, these are the related features (as historical sales) of similar (wrt sales behavior) products.

We evaluate our approach on the real data collected by food wholesaling and retailing company. The results demonstrate that (1) our ensemble learning approach can perform much better than the currently used baseline, (2) we can handle seasonal changes with ensemble learning better if feature set for a target product is complemented with features of related product (having similar sales pattern), and (3) an ensemble can become more accurate if information about the weather and holidays is presented explicitly in a feature set.

The rest of the paper is organized as follows. In Section 2 we discuss the problem of food sales prediction in more detail focusing on the Sligro company case. Section 3 gives a brief introduction to the problem of concept drift and provides a compact overview of the common approaches for handling concept drift. In Section 4 we outline the ideas of the dynamic integration of classifiers and motivate the

construction of additional features. The experimental results are discussed in Sections 5. We briefly conclude with a summary and discussion of the further work in Section 6.

## 2. Challenges of food sales prediction

In this section the describe a case of Sligro Food Group N.V., which encompasses food retail and food service companies selling directly and indirectly to the entire Dutch food and beverages market. The group pursues a multi-channel strategy, covering various forms of sales and distribution (cash-and-carry and delivery service) and using several different distribution channels (retail and wholesale). Sligro has about 60.000 products in stock.

In general, different kinds of the food sales predictions are required for performing different business operations. These kinds include first of all next day, next week, and next month predictions. Daily predictions based on a moving average over different nearest neighbors work reasonably well in this setting and wrong predictions can be often compensated by a human involvement. Weekly predictions are essential for wholesaling of food and food-related products and it is considered to be a more challenging and responsible activity. We focused on the weekly predictions due to this reason and also because of the fact that the weekly predictions are known to be the weakest point in the existing infrastructure.

Currently, moving average and simple regression models for baseline prediction are used and then managers use their domain expertise (knowledge about how products are doing during this time of the year, good or bad weather, school holiday period, promotion and advertisements possibilities, etc) to make the adjustment for the baseline predictions.

Predictions based on moving averages may work reasonable well when demand is close to level. When demand follows trend or seasonal patterns these methods do not perform well due to the slow reaction to the changes. A typical case of such behavior can be seen from Figure 1 (top). Each time the sales start to rise the prediction rises a couple of weeks later and vice versa when demand decreases. The most important points in the pattern are the increase at the beginning of the season and the decrease at the end of the season which can be described as "when to start accelerating or when to start pulling the brakes". In the first case, stock often occurs since not as much is ordered as sold whereas in the second cases the products expiration date expires as a result of ordering too much when sales are already decreasing. Another case (bottom plot of Figure 1) shows that moving average just filters most of the changes in the demand as it was noise.

These typical drawbacks of the moving average method is the main reasons that predicting demand for a product having some seasonal patterns correctly is mostly done with
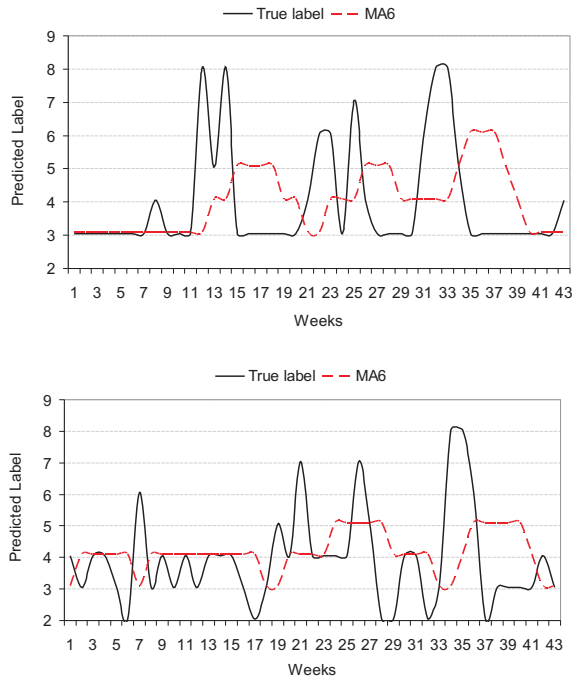
135

**Figure 1. Predicting with moving average as a baseline.**

the human expertise. Managers often try to improve the performance in these periods by increasing the number of safety days which results in a higher amount of products to be ordered. However, this method only works in case the managers are experienced enough and even then it is still difficult to predict at which point in time this should be done. Another typical approach is to have a number of triggers and reminders that should hint about the coming (school, national or religious) holidays, (warm or cold, sunny or rainy) weather and other factors that influence demand of particular food products. However, human factors like frustration, overload of information, lacking expertise (especially with a new personnel or for a new set of products), and simply forgetfulness may result in mistaken predictions and poor decision making.

In the rest of this paper we will study whether it is possible to develop a data mining approach for providing a better baseline for human experts, whether human domain expertise can be presented (either in explicit or implicit form) as an input to the data miming approach for sales prediction.

Certainly, besides intelligent (whether data-driven, knowledge-driven or mixed) forecast generation tools and techniques, (right) data availability, training in forecasting, and other organizational aspects have tremendous impacts on the food sales forecasting process.

Although the current state of our research does not allow to concentrate on the usability and utility issues, after introducing our ensemble learning approach for food sales prediction, we will particularly concentrate on the issue of data availability, trying to demonstrate that besides the past history of demand for a product that is essential in the preparation of seasonal patterns, it is equally important to consider contextual information (promotion campaigns, multichannel advertisements, etc) and other sources of information, including weather history and forecasts (temperature, rainfall, hours of sunshine etc.), civil (national and school holidays) and religious (holidays and fastening) calendars, and possibly other kind of information that may impact the demand. But first, we overview the problem known in data mining as concept drift with some more details.

## 3. Concept drift and seasonal changes

A difficult problem with learning in many real-world domains is that the concept of interest may depend on some hidden context, not given explicitly in the form of predictive features. Changes in the hidden context can induce more or less radical changes in the target concept (thus, consumers' preferences, habits, and interests which effect the demand and some of which change over time, are often not possible to observe directly), which is generally known as concept drift [23]. Popular examples of concept drift include antibiotic resistance, where pathogen sensitivity may change over time as new pathogen strains develop resistance to antibiotics that were previously effective, electricity demand prediction and others. Under the assumption of concept drift, an effective learner should be able to track changes and to adapt to them as quickly as possible.

Changes in hidden context may not only be a cause of a change of target concept, but may also cause a change of the underlying data distribution. Even if the target concept remains the same, and it is only the data distribution that changes, this may often lead to the necessity of revising the current model, as the model's error may no longer be acceptable with the new data distribution. The need to the change of current model due to the change of data distribution is called virtual concept drift [22]. Virtual and real drifts often occur together. From the practical point of view it is not important, what kind of concept drift occurs, real or virtual, or both. In all cases the current model needs to be changed. Three approaches to handling concept drift can be distinguished: (1) instance selection; (2) instance weighting; and (3) ensemble learning [23]. Other categorization of the different approaches is based on whether they use any detection mechanism (changes in the distribution or drop of accuracy) or handle this implicitly [4].

In instance selection, the goal is to select instances relevant to the current concept. The most common concept

drift handling technique is based on instance selection and consists in generalizing from a window that moves over recently arrived instances and uses the learnt concepts for prediction only in the immediate future [23]. Many case-base editing strategies in case-based reasoning that delete noisy, irrelevant and redundant cases are also a form of instance selection [2].

Instance weighting uses the ability of some learning algorithms such as Support Vector Machines (SVMs) to process weighted instances [6]. Instances can be weighted according to their "age", and their competence with regard to the current concept. Klinkenberg [6] demonstrates in his experiments that instance weighting techniques handle concept drift worse than analogous instance selection techniques, which is probably due to overfitting the data.

Ensemble learning maintains a set of concept descriptions, predictions of which are combined using a form of voting, or the most relevant description is selected. Street and Kim [14] and Wang *et al.* [21] suggest that simply dividing the data into sequential blocks of fixed size and building an ensemble on them may be effective for handling concept drift. Stanley [13] and Kolter and Maloof [7] build ensembles of incremental learners in an online setting, starting to learn new base classifiers after fixed intervals, while continuing to update the existing ones. All incremental ensemble approaches use some criteria to dynamically delete, reactivate, or create new ensemble members, which are normally based on the base models' consistency with the current data.

Ensemble learning is among the most popular and effective approaches to handle concept drift, in which a set of concept descriptions built over different time intervals is maintained, predictions of which are combined using a form of voting, or the most relevant description is selected [8]. Interestingly, that in time series research it was also recently pointed out [24] that no single method is best in every situation and that combining different models is an effective and efficient way to improve accuracy of (sales) prediction.

## 4. Our approach for food sales prediction

### 4.1. Ensemble learning approach with dynamic integration of classifiers

A number of weighing and selection techniques have also been proposed to address the task of integration. One of the most popular and simplest selection techniques is Cross-Validation Majority (CVM) [12]. In CVM, cross-validation accuracy for each base classifier is estimated, and the classifier with the highest accuracy is selected. At present, the most common integration approach with ensembles for handling concept drift is weighted voting, where each base classifier receives a weight proportional to its relevance to the

current concept. With weighted voting, lower weights can be assigned to predictions from base classifiers simply because their global accuracy on the current block of data falls, even if they are still good experts in the stable parts of the data.

In the real world, concept drift may often be local, for example in multi-confessional countries only particular subgroup of people will be fastening during particular periods in their religious calendar (which is often not present as predictive features), while most of the eating habits of other subgroups may remain the same. This is an important problem with most of the existing ensemble approaches for handling concept drift. One solution to this problem is the use of dynamic integration of classifiers, in which the models are integrated at an instance level according to their local accuracies [17].

The idea of the use of dynamic integration for handling concept drift was introduced in [16] with experiments focusing on the problem of antibiotic resistance in nosocomial infections, and in [17] the dynamic integration approach is presented in the level of detail necessary for the possible implementation. Here, we only highlight the key ideas.

In dynamic integration, each base classifier receives a weight proportional to its local accuracy in the neighbourhood of the current test instance, instead of using global classification accuracy as in normal weighted voting.

In dynamic integration information about each new instance is taken into account. We consider in our experiments three dynamic techniques based on the same local performance estimates (of each base classifier for each instance of the validation): Model Selection (MS), Model Voting (MV), and Model Weighed Voting (MWV).

Testing a new instance begins with determining $k$ nearest neighbours for this instance from the validation set. Then, weighted nearest neighbour learning is used to predict the local performance of each base classifier. Different distance functions can be used for determining the neighbourhood of the current test instance in dynamic integration. The simplest and most common way is to use the heterogeneous Euclidean/overlap metric (HEOM) in the instance space so that the Euclidean distance is used with numeric features, and the overlap distance with categorical features in order to find a distance between two instances. A simple weighting function that just raises the distance to a negative power is perhaps the most commonly used weighting function in locally weighted learning, and is often used for the Euclidean and overlap metrics.

Then, MS simply selects a classifier with the best local predictive performance. In MV, from each neighbor a classifier with the best local predictive performance is selected, and then the final classification is produced using voting. In MWV, each base classifier receives a weight that is proportional to its estimated local performance, and the final

classification is produced using weighted voting (of all classifiers from each neighbor).

Dynamic integration was successfully applied in a number of contexts with stationary problems, outperforming other integration methods. In [18] the three dynamic integration techniques were used to combine the base classifiers generated with bagging and boosting, improving the predictive performance of the two most popular ensemble techniques. In [15] dynamic integration was considered in the context of ensembles with base classifiers generated on different feature subsets (using so-called ensemble feature selection). In [11] an adaptation of the dynamic integration techniques to regression is considered and applied for ensembles generated using the random subspace method.

In the context of ensembles for handling concept drift the most commonly used integration techniques are voting and weighted voting, although as it is demonstrated in some recent research [17][4] they are not always the most appropriate techniques.

## 4.2. Data preprocessing and feature space construction

Sligro maintain data warehouse where raw data can be accumulated and aggregated. For the purpose of weekly sales predictions, the original data is first transformed using Symbolic Aggregate Approximation (SAX) [9] into the stings of symbols. We used the alphabet of 8 symbols that reflect the variation in sales from very low (1) to very high (8). Different experimental studies suggested that SAX may improve the results of time series clustering, classification and novelty detection [9].

Besides the 23 standard features (including week, moving averages, slopes, and history of sales), we introduced 60 external features that we believed may be helpful in sales prediction.

Business analysis has shown that some correlation with sales figures can be found in features describing holidays, school holidays and weather properties. Changes in those domains have been identified as a trigger for an increase (or drop) in sales. We have added features from the KNMI (Koninklijk Nederland Meterologisch Instituut) and the MinOCW (Ministerie van Onderwijs, Cultuur en Wetenschap) to describe these contexts. Respectively, the average temperature per week, the holidays (15 in total) and school holidays per week have been added as additional features. Through a year, Sligro has several periods per year in which promotions are held to sell more of a specific products. We have extracted promotions from the database of the wholesaler and added one feature to describe whether a product was promoted in that week or not. Each of the introduced features is added for the current week, the next week and the week after that due to a required delay between the actual buying of the products and the consumption in the week of the holiday.

Although we learn an ensemble for each product independently, many products are certainly not completely independent. Therefore, intuitively, it might be logical to assume that if we use information on sales of the related products this will help us to learn some patterns of the sales of one product triggering the sales of the second product.

We employed agglomerative hierarchical clustering (AHC) to group together products with a similar sales pattern.

Since we can expect that there is not necessarily a one-to-one, linear correspondence of sales behavior of related products, we can use a distance measure that allows non-linear mapping of time series. The Dynamic Time Warping (DTW) distance measure is one of such measures that allows a non-linear mapping of one signal to another by minimizing the distance between the two. DTW has long been known in speech recognition community and more recently was introduced and successfully applied in DM community as a utility for various tasks in time series mining including classification, clustering, and anomaly detection, especially when time series are transformed into a symbolic form [9]. Besides DTW, we tried other popular distances including Euclidean distance, Longest Common Subsequence (LCS) [20], and Compression-based similarity (CDM) [5]. Euclidian distance performed rather poorly as expected. DTW and LCS resulted in rather similar dendrograms. CDM produced less intuitive results, likely due to the rather short length of time series.

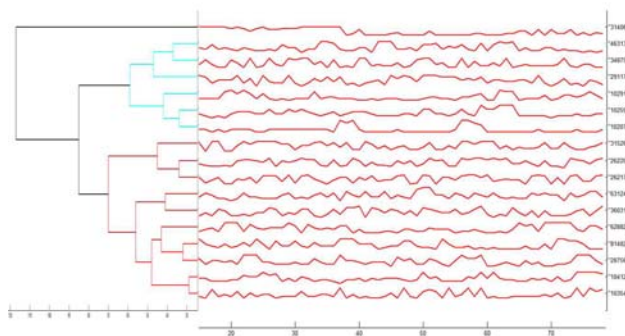Figure 2 shows the dendrogram for the dataset we have chosen to conduct experiments with.



**Figure 2. AHC of selected products; complete linkage, DTW distance.**

We added the six historical weeks of each related time series as this may help in finding delays in two series, for example, the sales of one product triggering the sales of the

138

second product. The number of related product to be taken into account is defined according to the dendrogram cut. In our case it has been 3-5 related products. The complete feature construction process is presented in Figure 3.
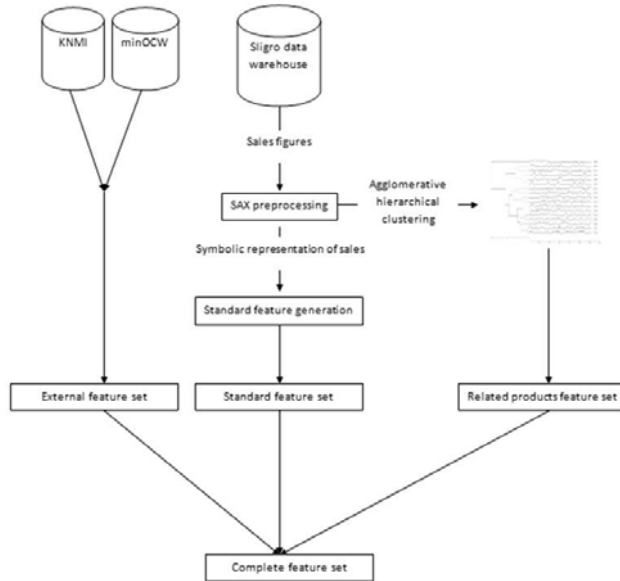


**Figure 3. Feature construction process.**

However, we can expect that there will be a trade-off between extra information available in all these contextual, external and related products' features and the effect of the curse of dimensionality. Having about 50-60 instances to learn an ensemble and about 100 features that describe each instance clearly suggests the necessity of applying some feature selection or feature extraction. We leave the thorough analysis of this issue for the further study. Here, we just employ a simple filter-based individual feature selection and analyse the performance of ensembles on different groups of features with and without feature selection.

## 4.3. Experimental settings

In our experimental studies we used an implementation based on the RapidMiner library formerly known as YALE [10] (available at http://rapid-i.com/). Default settings were used in the learning algorithms for our experiments.

**Target data selection.** 17 datasets of sales figures were extracted from the data warehouse. Each dataset consists of about two years of sales data. Each example represents the number of items sold per week.

These 17 products are selected based on the amount of sales, i.e. they are the best selling products in a group and are therefore, when regarding sales amounts, relatively important.

The selected products are quite similar according to their properties, i.e. are "semantically" related. Although we do not study this aspect in detail, such selection is also in line with our expectation of finding relevant information from related products (with regard to their sales figures).

A priori it was not known whether this subset of products would be responsive to the features added to the dataset. However, business analysis has shown that these product groups are known to contain products that change over the season. Thus, these selected products are also suitable for validating our approach focused on better handling of seasonal changes and the improvement of prediction accuracy.

**Ensemble learning.** For each product data was divided into the training set and in the test set having 52 and 43 instances correspondingly.

In the learning phase, the training set, consisting of 52 instances, is portioned into folds. Cross validation is used to estimate the errors on each instance of all base learners on the training set. Next the meta-level training set is formed. It contains the same features as were used in the training set and in addition to that, the performance matrix that contains the errors for all base learners. In the application phase, the combining classifier is used to classify a new instance.

In this work we use a pool of 24 heterogeneous classifiers which are retrained after each window shift. Heterogeneous ensemble consists of C4.5 and CHAID decision tree learners, two rule learners (with and without pruning), 1-nearest neighbor and attribute base votes lazy learners, and two functional learners, LibSVM and regression. Each learner was learnt on the three different windows; last 52 weeks, last 26 weeks, and last 13 weeks.

We experimented with 5 different sizes of neighbourhood $k$; 1, 3, 7, and 15. We have not tried larger $k$'s since accuracy normally decreases with the increase in the size of neighbourhood, becoming closer to static voting [17].

We evaluated three dynamic integration strategies described above; MS, MV, and MWV.

## 5. Experimental results

The analysis of the performance of the different dynamic integration strategies showed that MS performed worse than MWV and MV in most of the cases, MWV and MV often produced rather similar results with small $k$, yet MWV was slightly ahead of MV in most of the cases. Therefore, we will present the other results for MWV only.

Dynamic integration was not very sensitive to the size of neighbourhood. A reason for that is the locally weighted

learning scheme used, with which the more distant an instance is from the current test instance, the less influence it will have on the prediction of local performance.

In Figure 4 the average MSE (mean squared error) for the worst case scenario (WC), moving average over 6 weeks (MA6), and different ensembles are presented. Different ensembles were built on the different groups of features; original set of "standard" features (O), ensembles built on original plus external feature set (holidays, weather, promotions) with and without feature selection (OE/FS and OE respectively), ensembles built on original plus features from the related products (OR/FS and OR),and ensembles built on the complete set of all these feature (OER/FS and OER). Since we can estimate the error for the worst case scenario (all the true labels are known) we can scale MSEs to the corresponding accuracy estimates of ensemble performance (Figure 5).
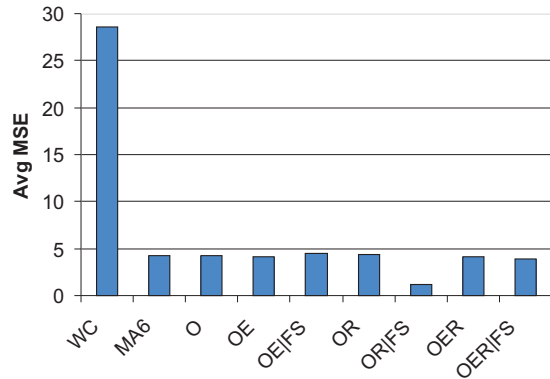


**Figure 5. Average accuracies of different approaches in correspondence with MSE results in Figure 4.**



**Figure 4. Average MSE for the worst case scenario (WC), moving average over 6 weeks (MA6), and different ensembles.**

**Table 1. Win/tie/loss pairwise comparison of different approaches.**

| w/t/l | O | OE | OE\|FS | OR | OR\|FS | OER | OER\|FS | MA6 | Total |
|---|---|---|---|---|---|---|---|---|---|
| O | | 5/7/5 | 5/5/7 | 3/7/7 | 16/0/1 | 7/7/3 | 6/8/3 | 5/7/5 | 47/41/31 |
| OE | 5/7/5 | | 2/8/7 | 4/8/5 | 16/1/0 | 5/11/1 | 5/8/4 | 7/5/5 | 44/48/27 |
| OE\|FS | 7/5/5 | 7/8/2 | | 5/7/5 | 16/1/0 | 10/3/4 | 9/6/2 | 6/8/3 | 60/38/21 |
| OR | 7/7/3 | 5/8/4 | 5/7/5 | | 16/1/0 | 7/7/3 | 6/9/2 | 3/11/3 | 49/50/20 |
| OR\|FS | 1/0/16 | 0/1/16 | 0/1/16 | 0/1/16 | | 1/1/15 | 0/1/16 | 1/0/16 | 3/5/111 |
| OER | 3/7/7 | 1/11/5 | 4/3/10 | 3/7/7 | 15/1/1 | | 4/9/4 | 6/4/7 | 36/42/41 |
| OER\|FS | 3/8/6 | 4/8/5 | 2/6/9 | 2/9/6 | 16/1/0 | 4/9/4 | | 5/10/2 | 36/48/35 |
| MA6 | 5/7/5 | 5/5/7 | 3/8/6 | 3/11/3 | 16/0/1 | 7/4/6 | 2/10/5 | | 41/45/33 |

It can be seen from the figures than on average different ensemble learning approaches (except OR/FS) are not doing much better than the moving average. Only ensembles built on the feature sets constructed with the standard set and (selected) features adopted from the related products resulted in about 10% accuracy gain on average. This clearly suggests that 1) the idea of using features from the related products is feasible and 2) feature construction/selection process in not trivial at all and need to be investigated further. Since averages hide a lot of potentially interesting information, we made a pairwise comparison of these approaches for each product. The summary is presented in the table below.

Each cell in the table tells how many wins (increase in accuracy for at least 2%), ties, and losses (decrease in accuracy for at least 2%) one approach had against another.
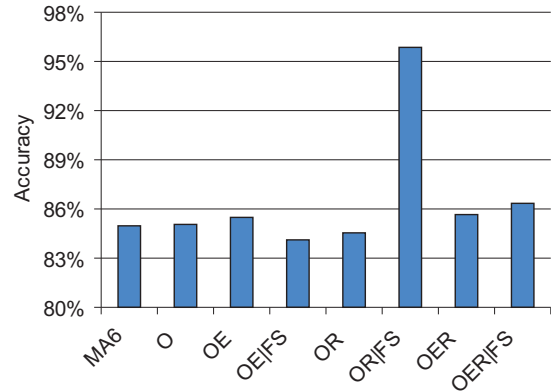
Thus, e.g. moving average had 5 wins, 7 ties and 5 losses against ensembles built on the standard feature set. The last column present the sums of all wins/ties/losses for each approach. It can be seen from the table that OR/FS is indeed a clear winner. It can be also seen that there are still many cases when ensemble learning performs better than moving average. However, FS does not seem to have a clear positive effect on different ensembles.

We studied the results of feature selection a bit further and found that, globally, most of the external feature were not informative and only about 5% (3 out of 60) of external features were selected on average for each product, in most cases - features presenting temperatures and promotions. This suggests that more laborious feature construction is necessary and that local dynamic feature selection might be helpful.

140

This was not the case with historical sales features from the related products, about 50% (14 out of 20) of external features were selected on average for each product. Surprisingly, more often the sales history from the two last weeks was less useful than the sales history from the 6-3 last weeks.

Since the experiments clearly showed that historical sales features from the related products allow to learn more accurate ensembles of classifiers, we were eager to get an idea regarding the tradeoff of getting extra information from the additional features and coping with the curse of dimensionality.

Figure 6 shows how MSE changes with adding more and more features from one, two or three related products (according to DTW distance). Filter feature selection was applied on the added feature subset in each case. It can be seen from the figure that for many products (yet with some exceptions), the error increases with adding extra (potentially discriminative) features. Therefore finding a "right" related product is an important task.
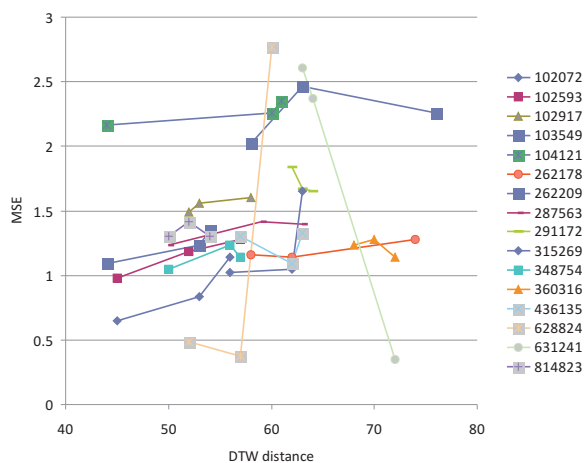


**Figure 6. Ensembles performance wrt added feature from the related products.**

A more detailed comparison of different approaches for two illustrative products is presented in Figures 7 and 8 respectively. It can be seen from the Figure 7 that the moving average is the closest to the worst case scenario. Ensembles learnt with related (Rel) or external(Ext) features added to the standard set perform rather similarly reacting to the changes reasonable well. Still, ensembles learnt on the whole feature set are more stable and accurate.

In Figure 8 the situation is rather different. Moving average performs quite well, ensembles learnt with external feature added to the standard set perform rather similarly

to the moving average, while ensembles learnt with related feature added to the standard set perform extremely poor. However, in this case ensembles learnt on the whole feature set were the most stable and accurate again.
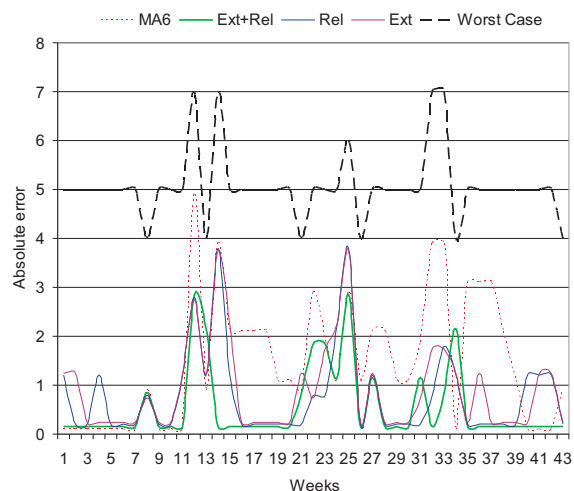


**Figure 7. Absolute errors in predictions with moving average (MA6) and different ensembles (product A).**

## 6. Conclusions and further work

Food sales prediction is an important and challenging problem having some connections to the problem of concept drift. In this paper we presented ensemble learning approach with the dynamic integration of classifiers for handling seasonal changes in sales.

Our experimental study demonstrated that ensemble learning approach performs for the subset of products much better than the currently used baseline. The results also showed that we can handle seasonal changes with ensemble learning for many products better if feature set for a target product is complemented with features of the most related products (having a similar sales pattern), and that an ensemble can become more accurate if information about the weather and holidays is presented explicitly in a feature set. The results demonstrate also the importance and necessity of the representation space improvement which otherwise may have a dramatic impact on the performance of ensembles.

We see the main novelty of this work in developing the elements of proactive handling of concept drift. Intuitively, if we observe a drift in sales for one product we can do a better prediction of sales for another related prod-
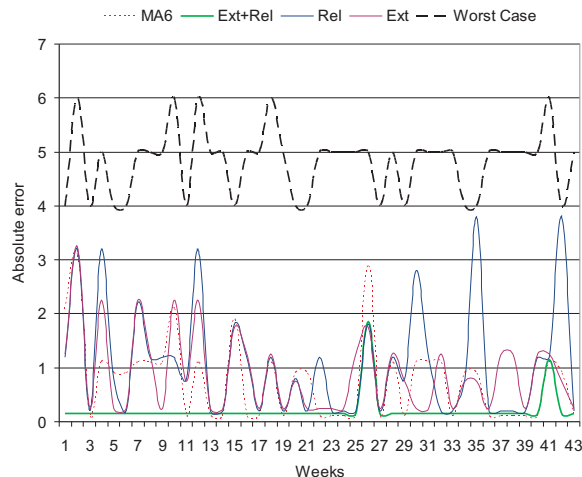
141

**Figure 8. Absolute errors in predictions with moving average (MA6) and different ensembles (product B).**

uct. In this paper we studied this idea only with ensemble learning approach and implicit use of information about the changes observed in related products in base classifiers. However, our further studies include applying similar idea of proactive handling of concept drift with other windowing and instance selection approaches for handling concept drift, which explicitly detect a change point. Although the preliminary results are promising, there is still much work ahead on validating the finding with larger set of products and extensive simulation studies on the artificial data. In this paper we estimated the similarity of different products based of their sales figure, i.e. time series matching. Our further work includes also applying market basket analysis for finding relations between the products.

Other directions of our further work in food wholesales prediction include studying the tradeoff of adding extra information from highly related time series vs. increase of dimensionality (under the limited number of data points for training), analyzing the benefits of ensemble learning with associative classifiers for knowledge base development; developing different evaluation strategies for (changes in) sales prediction; analyzing the performance of different ensemble learning approaches for making longer term prediction; and improving weighted voting procedures by adjusting competencies of classifiers with respect to the recent sales behavior of related products.

Last but not least we study the different ways of measuring the sales prediction accuracy that is essential for monitoring and comparing the performance of employed approached. Our experience shows that it might not be appropriate to use one global measure for all products within a business as a result of rather different behavior in sales, volume and supply characteristics. Although a few similar problems have been studied for example in learning from imbalanced datasets and cost-sensitive learning, here we may need to come up with some new utility definitions.

## References

[1] D. Adebanjo and R. Mann. Identifying problems in forecasting consumer demand in the fast moving consumer goods sector. *Benchmarking: An International Journal*, 7(3):223–230, 2000.

[2] S. J. Delany, P. Cunningham, A. Tsymbal, and L. Coyle. A case-based technique for tracking concept drift in spam filtering. pages 3–16, Queens' College, Cambridge, UK, 2004. Springer, Springer.

[3] P. Doganis, A. Alexandridis, P. Patrinos, and H. Sarimveis. Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering*, 75(2):196–204, 2006.

[4] J. Gama and G. Castillo. Learning with local drift detection. In X. Li, O. R. Zaïane, and Z. Li, editors, *ADMA*, volume 4093 of *Lecture Notes in Computer Science*, pages 42–55. Springer, 2006.

[5] E. J. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S.-H. Lee, and J. Handley. Compression-based data mining of sequential data. *Data Min. Knowl. Discov.*, 14(1):99–129, 2007.

[6] R. Klinkenberg. Learning drifting concepts: Example selection vs. example weighting. *Intelligent Data Analysis*, 8(3):281–300, 2004.

[7] J. Z. Kolter and M. A. Maloof. Dynamic weighted majority: A new ensemble method for tracking concept drift. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, pages 123–130, Washington, DC, USA, 2003. IEEE Computer Society.

[8] L. I. Kuncheva. Classifier ensembles for changing environments. In F. Roli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems*, volume 3077 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2004.

[9] 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.

[10] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: rapid prototyping for complex data mining tasks. In T. Eliassi-Rad, L. H. Ungar, M. Craven, and D. Gunopulos, editors, *KDD*, pages 935–940. ACM, 2006.

[11] N. Rooney, D. W. Patterson, S. S. Anand, and A. Tsymbal. Dynamic integration of regression models. In F. Roli, J. Kittler, and T. Windeatt, editors, *Multiple Classifier Systems, 5th International Workshop, MCS 2004, Cagliari, Italy, June 9-11, 2004, Proceedings*, volume 3077 of *Lecture Notes in Computer Science*, pages 164–173. Springer, 2004.

[12] C. Schaffer. Technical note: Selecting a classification method by cross-validation. *Machine Learning*, 13(1):135–143, 1993.

[13] K. Stanley. Learning concept drift with a committee of decision trees, 2001.

[14] W. N. Street and Y. Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 377–382, New York, NY, USA, 2001. ACM.

[15] A. Tsymbal, M. Pechenizkiy, and P. Cunningham. Sequential genetic search for ensemble feature selection. In L. P. Kaelbling and A. Saffiotti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, pages 877–882. Professional Book Center, 2005.

[16] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections. In *Proceedings of CBMS 2006, International Symposium on Computer-Based Medical Systems*, pages 679–684, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[17] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Dynamic integration of classifiers for handling concept drift. *Information Fusion*, 9(1):56–68, 2008.

[18] A. Tsymbal and S. Puuronen. Bagging and boosting with dynamic integration of classifiers. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 116–125, London, UK, 2000. Springer-Verlag.

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

[20] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. J. Keogh. Indexing multidimensional time-series. *VLDB J.*, 15(1):1–20, 2006.

[21] H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 226–235, New York, NY, USA, 2003. ACM.

[22] G. Widmer and M. Kubat. Effective learning in dynamic environments by explicit context tracking. In *ECML '93: Proceedings of the European Conference on Machine Learning*, pages 227–243, London, UK, 1993. Springer-Verlag.

[23] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Mach. Learn.*, 23(1):69–101, 1996.

[24] P. G. Zhang. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159–175, 2003.