Modelling of the Complex Data Space

Architecture and use cases from NRG4CAST project

Klemen Kenda, Maja Škrjanc, Andrej Borštnik
AI Lab (E3)
Jozef Stafan Institute
Ljubljana, Slovenia
klemen.kenda@ijs.si

***Abstract*—The following contribution offers technical solution for heterogeneous multivariate data streaming modelling built on top of open-source QMiner platform. The presented infrastructure is able to receive data from different sources (sensors, weather, weather and other forecast, static properties …) with many different properties (frequency, update interval …), it is able to merge and resample this data and build models on top of it.**

Technology was used to prepare prediction models for 5 different energy related use cases which include public buildings, thermal plant production, university campus buildings, EPEX energy spot market prices and total traded energy. Average relative mean absolute error of the model predictions varies between 5-10%, and qualitative analysis of predictions shows significant correlation between predictions and true values.

Keywords—data fusion, modelling, prediction, data streams, sensor data, sensor networks, model trees, Hoeffding trees, SVM

I. INTRODUCTION

The paradigm of modelling has adjusted to the nature of most of the data nowadays – the data is coming in a continuous stream. Additionally there are a lot of open data available in the World Wide Web, which can significantly contribute to the accuracy of prediction models. Many of the classical prediction methods have been already ported to the streaming scenario, however – one can find a demanding technical challenge in the demand for fusion of heterogeneous multivariate data sources to prepare valid feature vectors for modelling.

In the paper we are addressing methods for predicting energy-related phenomena in public buildings, energy markets and energy provides. The contribution of this paper is three-fold.

Firstly we are examining the potential additional data sources for the problem in question and are suggesting a full set of features to assist with successful modelling. Secondly we provide evaluation of different prediction methods and behavior of models with different feature sets. Thirdly – and most importantly – we are proposing a methodology and providing a prototype for fusion of heterogeneous multivariate data sources for modelling.

II. FEATURES AND FEATURE VECTORS

Prediction capabilities of the models are in general more dependent on the used features than on the modelling method selected. Extensive analysis of the use cases [1] has shown that the following types of features should be considered for modelling: sensor features, forecasts and static properties. Table I depicts an example of a full feature vector for modelling energy consumption of a National Technical University of Athens (NTUA) campus building.

Sensor data are streaming data data in the "classical" sense of the word. The system receives data in an orderly fashion. There are a few exceptions, though. Data is not coming as it is being generated. Often systems implement some sort of buffering (to avoid overhead, network congestions and similar) or there are just some technical issue preventing data to be received in a true on-line fashion. We need our system to deal with such exceptions.

Prediction data are different in the way that predictions can change through time. For example: weather forecast for a day after tomorrow will be refined tomorrow and different values will have to be taken into account. Many streaming mechanisms do not work in such a scenario. Also – the data we have is not aligned with the measurement, but usually extend to and beyond the prediction horizon.

Static properties data are concerning time of day, week, day of year, time of day, holidays, working days, weekends, moon phase etc. This is the data that can be pre-calculated and is usually pushed into the prediction engine at once – in the initial data push.

TABLE I. FULL FEATURE VECTOR FOR NTUA USE CASE

Type	**Feature**			
	Name	***UoM*[a]**	***Value*[b]**	***Aggr.*[c]**
Sensor	current_l1	A	X(0)	
	current_l2	A	X(0)	
	curent_l3	A	0	
	energy_a	kWh	0, -1h, -1d	
	demand_a	MW	0	yes

Type	Feature			
	Name	***UoM***[a]	***Value***[b]	***Aggr.***[c]
	demand_r	kvar	0	
Weather	temperature	°C		yes
	wind speed	m/s		yes
	wind direction	°		yes
	visibility	km		yes
	humidity	%		yes
	pressure	mbar		yes
	cloud cover	%		yes
Weather forecast	temperature	°C	t	
	wind speed	m/s	t	
	wind direction	°	t	
	cloud cover	%	t	
	humidity	%	t	
Static properties	weekday		t	
	dayOfWeek		t	
	month		t	
	working day		t	
	working hour		t	
	holiday		t	
	day before holiday		t	
	day after holiday		t	

[a.] Unit of measurement

[b.] Value, expressed with relative time (0 = current timestamp, -1h je timestamp 1 hour ago; t denotes the timestamp of prediction)

[c.] Configuration of aggregates is much more complex, further details can be found in [1]

Weather forecasts have been provided by Forecast.IO[1] web-service, weather data has been provided by OWM (Open Weather Map[2]).

III. HANDLING MULTI-MODAL DATA

Each type of data, explained in Section II requires special handling throughout the pipeline, which transports the data from the data source to the final model.

In our design we have used two types of stream processing engine instances (built on top of QMiner[2]) to handle this data: data instance (handles loading of the data and initial data fusion – calculating aggregates) and modelling instance (final data fusion and modelling). The setup used in the NRG4CAST project is depicted in Fig. 1. More detailed view of the components described below is depicted in Fig. 2. Description of components also explains all the steps needed for multi-modal multivariate data fusion on a streaming data sources for modelling.

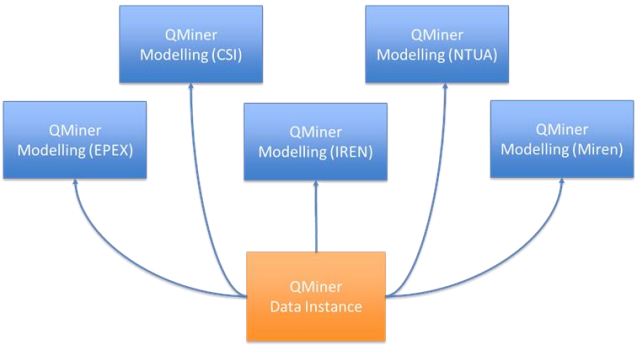

Fig. 1. Data and Modelling instances of QMiner in the NRG4CAST platform.

Data Instance includes the following components:

- **Push (time sync) Component**
  This component overcomes the problems, caused by unsynchronized arrival of sensor, forecast and static properties data. The component is invoked for a group of data streams (relevant for final model) arriving to the Data Instance. The components determines the lowest possible timestamp, where there are data in the Data Instance. Then it pushes items from all the streams and orders them by a timestamp. This makes it possible for the Modelling Instance to implement normal streaming algorithms on top of the received data. The pushed data includes measurement data as well as aggregates.

Modelling Instance includes the following components:

- **Store Generator**
  Modelling instance needs to provide stores for all the data it will be receiving and for all the merged data streams. This includes merged stores by the group of sensors and a meta-merged store with all the data.
- **Load manager**
  Load manager component is the one that invokes the Push Component. It provides push component with the list of relevant data streams and timestamp of the timestamp of the last received measurement. Load manager is loading the following data separately: sensors, properties, forecasts.
- **Receiver**
  Receiver listens to the data, sent by Push Component. Its sole purpose is to write the data in the appropriate stores. It also needs to take additional care that no record is duplicated.
- **Merger**
  Merger Component is a universal component that takes a group of data streams (these groups consist of one kind of streams – sensor data, properties, predictions) with arbitrary timestamps and joins all the measurements in a single store (table). Merger only

[1] http://forecast.io/
[2] http://openweathermap.org/

works with data items that do not break the timeline. Result of the merger is a huge table with data for each single timestamp in the source data.

- **Resampler**
  Merger data needs to be resampled to the relevant interval. In NRG4Cast this interval is mostly 1 hour. All the other measures are irrelevant. Different interpolation methods can be used to provide the relevant record (previous, linear). Records are written in a corresponding data strore.
- **Meta Merger**
  As the dynamics of the different group of data (sensor data, predictions and properties) is different, data is received at different times. Meta-merger provides a full data record
- **Semi-automated modeller**
  Modeller is described in more detail below.

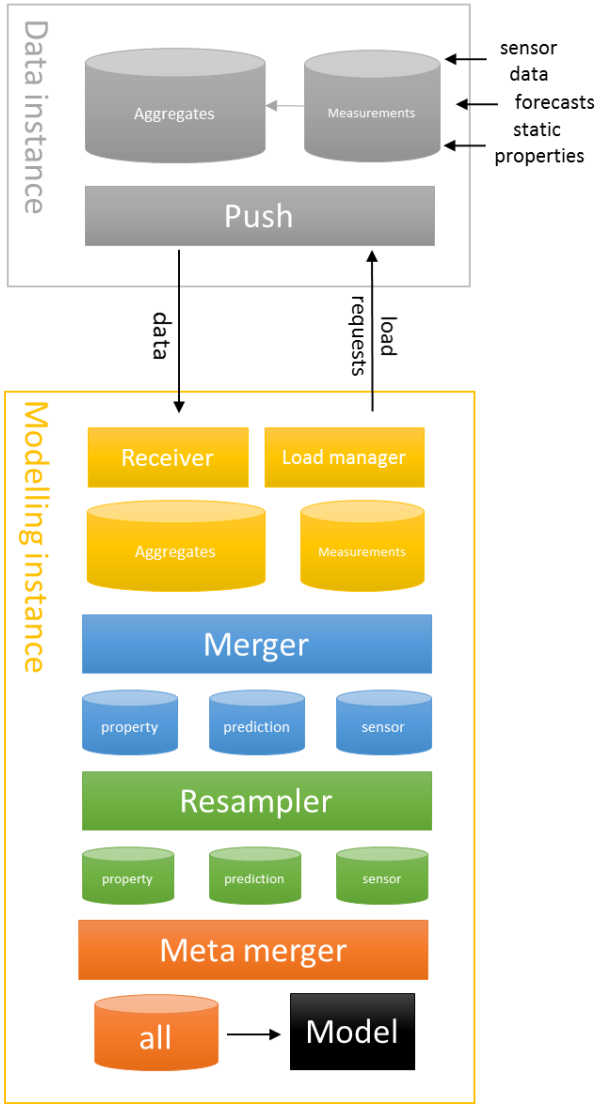

Fig. 2. Architecture to support multi-modal data fusion.

The nice property of such a system is, that it can also emulate the streaming setup from static data and can therefore seamlessly transform between classical and streaming data mining scenario.

IV. Modeller

The proposed methodology follows the on-line learning paradigm. All the evaluations were done on a real-time or simulated stream of real data.

Each of the components in the Modelling Instance from Fig. 2 needs to be configured for the specific task concerning a specific model. In most cases these configurations share the same content (relevant sensors). NRG4CAST platform has therefore been implemented with a single configuration structure that is able to automatize the data flow from the load manager to the modelling component. A technical description is available in [1].

The following methods have been implemented in the platform and experiments have been conducted to select the most appropriate methods for a particular use case:

- Linear regression (LR)
- Support Vector Machine Regression (SVMR)
- Ridge Regression (RR)
- neural networks (NN)
- moving average multiple models (MA)
- Hoefding trees (HT)

Most of the methods were adjusted to work in a streaming scenario, except SVMR, which used repetitive learning.

Properties of the predicted values have enabled us to use well known evaluation metrics. We have observed mean error (ME), root mean squared error (RMSE) and R^2 measure to determine the best possible model.

V. Results

A great number of experiments have been conducted to determine the best combination of prediction method, feature set and parameters that fit best to a scenario. A thorough result set for the following use cases can be seen in [1]:

- EPEX spot market energy price
- Turin public buildings
- Reggio Emilia thermal plant (IREN)
- Athens (NTUA) university campus buildingsž

Results from only Turin public buildings use case will be presented in this paper.

A. Turin pulic buildings results

There was 3.3 years of valid fused data (from June 2011 until October 2014) available for the experiments. We have used 2 years of data for learning and 1.3 years of data for evaluation. There was total 48 different features in the feature set (as shown in TABLE I.). Demanded granularity of predictions was 1 hour, prediction horizon was set from 12 to 36 hours. Predictions needed to be generated for one day ahead

before 12:00 each day. Feature to predict was building consumption without cooling system.

As the data has a strong daily period the first modelling decision was to build 24 models for the task – each predicting for a specific hour of a day.

TABLE II. shows selected results from the experiments. Symbols denoting feature sets, used in the table, are: AR (auto-regressive sensor features), S (other sensor features), F (forecasts), W (weather), P (static properties), ALL (full feature set).

One interesting observation was, that the weather data (current) never improved the accuracy of predictions. Also – weather data from available global web services seems to be somewhat noisy. Historic weather forecasts from the webservice used are very accurate, which means that some bias of the on-line predictions might be lost. Effects to the modelling have not been studied.

TABLE II. FULL FEATURE VECTOR FOR NTUA USE CASE

Method-feature set (parameters)	Error Measure		
	ME	***RMSE***	***R^2***
SVMR-ARFP (eps=0.015)	-2,74	16,50	0,84
SVMR-ARP (eps=0.05)	-2,51	17,23	0,83
LR-ARFP	-3,24	17,96	0,81
LR-ARP	-3,46	18,19	0,81
SVMR-ALL (eps=0.05)	-1,96	18,67	0,80
LR-ARSFP	-0,78	19,54	0,78
LR-ARSP	-0,81	19,74	0,77
NN-ALL (6,lr=0.02)	0,32	19,90	0,77
HT-ARSFP	-2,69	20,02	0,77
MA (7)	0,01	30,89	0,44

The best results were gained with SVM regression with auto-regressive sensor features, weather forecast and static properties. As a baseline method moving average over one week has been used.

Some weight analysis has been done on the linear regression model. Auto-regressive features (last three values and moving averages of 1 day and 1 week) were important. Aggregates like min, max and variance were not used by the liner regression. Other sensor features (building total consumption and data centre cooling consumption) and their aggregates were important as well.

Surprisingly – weather forecasts do seem less important. Among them cloud cover and humidity contribute most. There is not involvement of temperature or wind data. As expected – also static properties play an important role: weekend, day after holiday and holiday, day before holiday and working hours.

Example prediction model results are depicted in Fig. 3. Results vary from very good to satisfactory. We can also see, that some of the features can not be satisfactory explained by the current feature set (discrepancy on the first day in the figure).

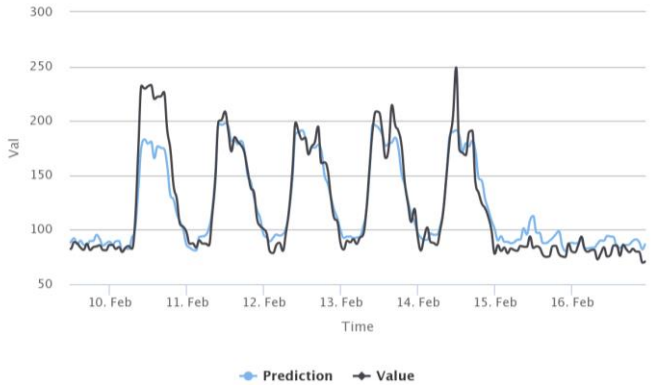

Fig. 3. Prediction for a selected Turin public building for a week in February 2015.

VI. CONCLUSIONS AND FUTURE WORK

In the paper we have presented results of the work done on preparation and evaluation of prediction models for energy-related scenarios. We have proposed a methodology of handling heterogeneous multivariate data sources in the modelling scenario and we provided a working prototype, which is running on a real live data stream.

We have presented an implementation of a particular model and the corresponding results on method selection. A comprehensive list of results can be found in [1].

Proposed methodology for data fusion (feature vector generation) has a potential to be widely exploited in many different scenarios. One example involving multi-level view over a complex sensor system has already been implemented and is presented in [3].

ACKNOWLEDGMENT

This work was supported by the ICT Programme of the EC under NRG4Cast (FP7-ICT-600074).

REFERENCES

[1] K. Kenda et al. “Modelling of the Complex Data Space”, NRG4CAST project deliverable D3.1, Ljubljana, November, 2014.

[2] B. Fortuna, J. Rupnik, J. Brank, C. Fortuna, V. Jovanoski and M. Karlovcec. “QMiner: Data Analytics Platform for Processing Streams of Structured and Unstructured Data”, Software Engineering for Machine Learning Workshop, Neural Information Processing Systems, 2014.

[3] K. Kenda, L. Stopar, M. Grobelnik. “Multilevel Approach to Sensor Streams Analysis”, Discovery Science, Bled, October, 2014.