Modelling in Energy Related Scenarios

Klemen Kenda
Jozef Stefan Institute
Jamova ulica 39
Ljubljana, Slovenia
klemen.kenda@ijs.si

Maja Škrjanc
Jozef Stefan Institute
Jamova ulica 39
Ljubljana, Slovenia
maja.skrjanc@ijs.si

Andrej Borštnik
Jozef Stefan Institute
Jamova ulica 39
Ljubljana, Slovenia
andrej-borstnik@hotmail.com

ABSTRACT
Fusing heterogeneous multivariate data in stream mining scenarios is a demanding task. Successful fusion requires a well-thought approach. We propose the use of a stream processing engine (SPE) that enables implementation of all the needed methods and ensures almost real-time responsiveness of the system.

In the paper we propose an infrastructure that is able to receive data from various heterogeneous sources (static properties, weather data and forecasts, other forecasts, and primarily sensor data). In the implementation of the proposed infrastructure we address issues related to the heterogeneous nature of the data, like different frequency, different update interval, and different nature of the data. The pipeline was used to prepare stream prediction models for five different energy-related use cases, which include public buildings, a thermal plant production, university campus buildings, and EPEX energy spot market prices alongside the total traded energy.

Keywords
Data fusion, modelling, prediction, data streams, sensor data, sensor networks, regression models, QMiner.

1. INTRODUCTION
Nature of obtaining data (wide availability, vast amounts) has changed the paradigm of modelling nowadays. It is fairly easy to measure certain phenomena with a continuous stream of measurements and it is even easier to add various open data to the set of modelling features.

Most of the systems are working in (almost) real time, which favours the streaming setup for modelling and predicting. Many of the classical prediction methods have already been ported to the streaming scenario. However in our work we have tackled a demanding technical challenge of fusing heterogeneous multivariate data sources to prepare valid feature vectors for modelling.

In this paper we are addressing methods for predicting energy-related phenomena in public buildings, energy markets, and at energy provides.

The paper presents an overview of potential additional data sources for the problem in question, showcases a suggested set of features for certain cases in energy related modelling, it provides an evaluation of different prediction methods, suggests an architecture for the multimodal stream modelling data fusion, and finally presents results from four different use cases processed within the platform.

2. FEATURES AND FEATURE VECTORS
Accuracy of prediction models is usually more dependent on the features used than on the modelling method chosen. Extensive analysis of five energy related use cases [2] has lead us to the following set of features with specific properties: sensor features, forecasts, and static properties. Table 1 depicts an example of a full feature vector for energy consumption modelling of the National Technical University of Athens (NTUA) campus building.

Table 1. Full feature set for campus building (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
	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	

Type	Feature			
	Name	*UoM*[a]	*Value*[b]	*Aggr.*[c]
	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]

The sensor data can be understood as the most fundamental streaming data. In an ideal case it is arriving to the SPE in (almost) real time as a conservative data stream (where measurements are ordered by a timestamp). Often transport systems implement different kinds of buffering, which means that the data is coming either with a delay, or even in chunks of multiple measurements. In a streaming scenario it is important that we are able to handle any exceptions and ensure that stream mining methods are fed feature vectors only when they include the most recent data.

Forecast data represents different kind of predictions, most commonly weather predictions. Forecasted data can also be classified as a stream, but with different properties. Forecasts get updated regularly. For example weather forecasts are updated every few hours and the system needs to be able to update the time series.

Static properties data is relatively easy to handle, as it can (in most cases) be calculated "a priori". Such data includes features like time of day, week, day of year, day of week, holidays, working days, weekends, moon phase etc. The data is similar to sensor data in the sense that it does not need to be updated and to prediction data in the sense that models usually refer to the future (and not current) values.

3. HANDLING MULTI-MODAL DATA

The implemented system has already been described in detail in [1]. In this contribution we will only describe the outline of the work done on the technical part and will rather focus on the results.

The system is built on top of the QMiner open-source platform [3]. We implemented two different systems: a data system and a modelling system. In the current setup we are running one instance of a data system (which collects data, orders it by time, handles properties and static data, and distributes it) and multiple instances of modelling systems (which merge separate data streams, resample them, create feature vectors, and model).

In the whole pipeline a number of components have been implemented: store generators, data adapters, aggregators, a time sync component, a load manager, a receiver, merger, a resampler, a meta-merger, and a semi-automated modeller.

The final result is a (single point) configurable system that is able to learn on the past data and give almost real-time predictions for various phenomena.

4. RESULTS

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

The following methods have been implemented in the platform:

- Linear regression (LR)
- Support Vector Machine Regression (SVMR)
- Neural networks (NN)
- Moving average multiple models (MA)
- Hoeffding trees (HT)

Most of the methods were adjusted to work in a stream mining scenario, except SVMR, which uses repetitive learning.

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

Feature sets in the results below are denoted by:

- AR – auto-regressive features
- F – weather forecasts
- P – static properties
- S – additional sensor data
- ALL –a full feature vector

Modelling demand in the energy related scenarios seemed to be quite unified for all the studied use cases. The customer is usually interested in an energy profile for the next 24-36 hour period at around 12:00 each day.

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

An interesting observation was that the weather data (current) never improved the accuracy of predictions. Weather data from available global web services also seems to contribute little to the prediction model. Historic weather forecasts from the web service used are very accurate (service provides the latest – short term - prediction for a location), which means that some bias of the longer term weather predictions might be lost. How this effects the modelling has not been studied.

4.1 Public Building (CSI)

Public building in Turin offered 2 years of data for the learning phase and 1.3 years of data for evaluation. The total number of features was 48. We were trying to predict building electricity consumption (cooling excluded).

All of the methods behaved quite well on this data set, but SVMR yielded the best results. All the methods in this case have been significantly better than the best base-line method (moving average over the last week). Results are shown in Table 2.

Table 2. Results from public building (CSI) use case

Method-feature set (parameters)	Error Measure		
	ME	*RMSE*	R^2
SVMR-ARFP (eps=0.015)	-2,74	16,50	0,84

Method-feature set (parameters)	Error Measure		
	ME	*RMSE*	R^2
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

In Figure 1 an example of prediction vs. measurements is depicted. This is a normal example form the validation part of the data set. We can see that the model is unexpectedly good with an exception of Mondays, where something that could not be modelled by the feature set appeared.

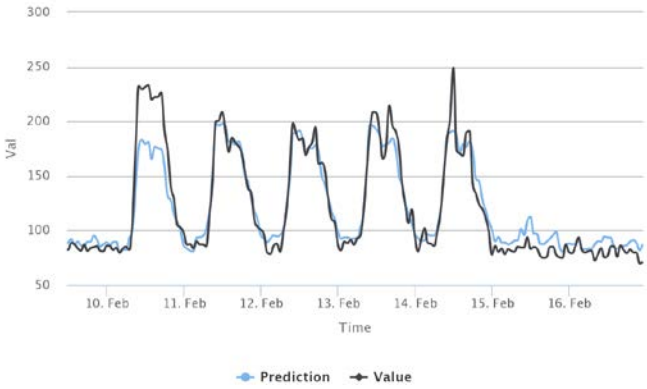

Figure 1. Prediction for a selected Turin public building, for a week in February 2015.

Further drill-down of the weights of the LR model has shown that the most significant features were the 1(one) week aggregates of auto-regressive and other sensor features. From the weather forecast feature temperature was surprisingly not among the most significant features, but cloud cover (solar radiation) and humidity were. Additional features such as day/hour classification (weekend, holiday, day after holiday, working hours, and heating season) were utilized the most.

4.2 University Campus Building (NTUA)

University campus of NTUA offers 5 years of valid data, which was divided into 3 years for learning and 2 years for evaluation. We are modelling average power demand for a selected building.

Results of the tests on this dataset have shown that the features provided for modelling are unable to handle all the dynamics of the system. Parts of the test data have been modelled quite well (see Figure 2), but the model did not handle the other parts well (see Figure 3). This might be a good indicator of possibly faulty, or simply missing data in the feature set.

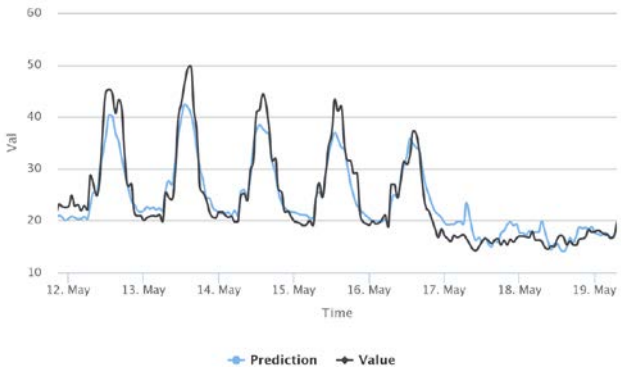

Figure 2. Model works well at some point.

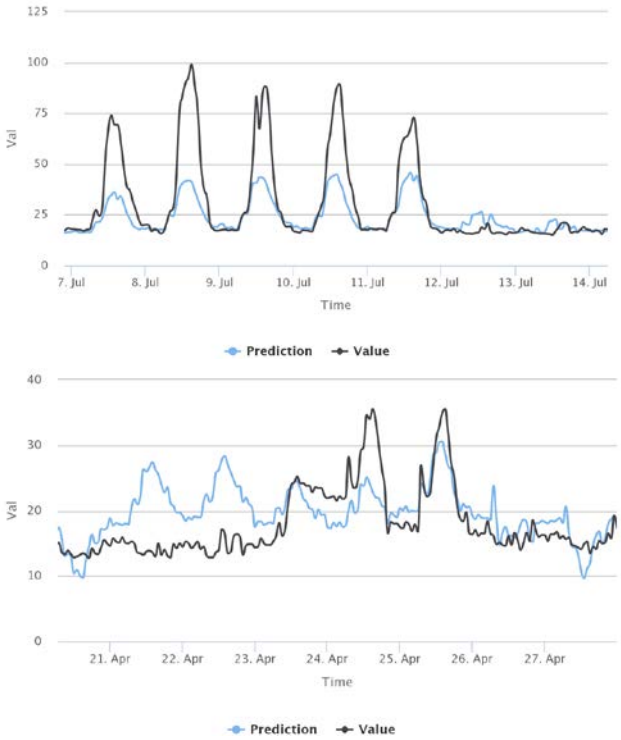

Figure 3. Unhandled exceptions in the modelling.

4.3 Energy Prices in Energy Spot Market (EPEX)

Data for the energy spot market has been scraped from the EPEX spot market web pages and streamed into the platform. At the testing phase there were 3 years of data available for learning and 1.4 years for evaluation. Energy prices and total trading energy for Germany were used in the experiments.

Energy prices strongly depend on the production of energy from the alternative energy sources. The production costs for such energy are usually very low or equal to zero, but the energy grid is not yet prepared for such irregular intake of energy. Excessive production of energy form alternate sources therefore results in lowered prices (sometimes even negative prices).

Feature vectors have therefore included data from 6 different weather stations across Germany, especially the wind data (speed and bearing) and cloud cover were expected to be the most important features.

Table 3. Results from energy prices (EPEX) use case

Method-feature set (parameters)	Error Measure		
	ME	***RMSE***	***R^2***
LR-ARSFP	-0,53	8,59	0,71
LR-ALL	-0,28	8,64	0,70
SVMR-ALL (c=0,037, eps=0,034)	1,01	8,94	0,63
LR-ARFS	-0,22	10,29	0,58
HT-ARSFP	-2,29	13,41	0,29

According to Table 3 the safest methods behave best. Weight analysis of the LR showed that the most important features were energy prices in the previous days, total traded energy averages for 1 week, 1 month, and minimum/maximum total traded energies for previous week.

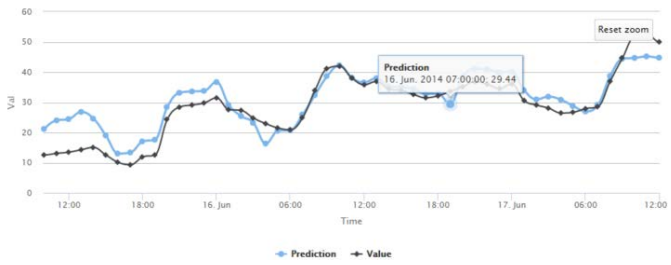

Figure 4. Prediction for EPEX use case for energy prices.

From the weather data it was interesting to see that wind bearing was the most dominant feature (as it was much more weighted than wind speed). Cloud cover has not contributed significantly to the behaviour of the models.

4.4 Thermal Plant (IREN)

1.6 years of data for thermal plant in Reggio nell'Emilia were available. 1.1 year was used in the learning phase and 0.5 years for testing. There were 43 features in the dataset.

During the experiments we were unable to satisfactory model part of the data and therefore some of the measures in Table 4 are distorted. The results for most of the data set are however very good, as can be seen in Figure 5.

Table 4. Results from thermal plant (IREN) use case

Method-feature set (parameters)	Error Measure		
	ME	***RMSE***	***R^2***
LR-ALL	-1,27	17,41	0,80
LR-AR	-0,08	17,94	0,79
MA (4)	-0,70	17,99	0,79
NN (4-6-3, lr=0.04)	-0,10	18,65	0,77
SVMR (c=0.03, e=0.02)	0,19	19,25	0,75

The weight analysis of the LR model shows significant contributions from most of the features.

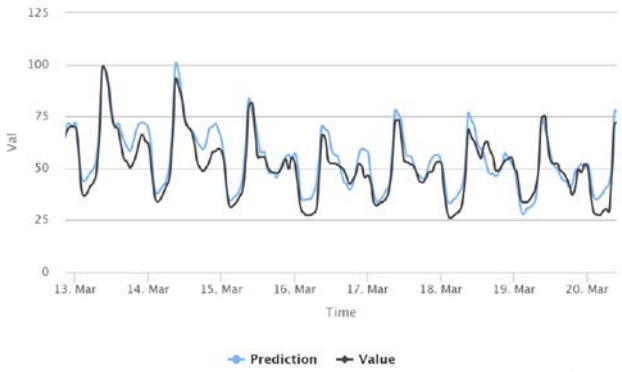

Figure 5. IREN use case prediction example.

5. CONCLUSIONS

In this paper we have presented models developed in the energy related scenarios using stream mining methods. We have developed a stack of components that are able to handle sensor data, forecasts, and static properties in a stream mining scenario. The platform has enabled us to provide streaming models for different energy-related phenomena. With the platform we are able to handle heterogeneous data from different independent data sources, such as different sensor systems, web services, or static flat files.

Accuracy of the developed models is mostly very good; however there are periods in the evaluation sets that we were sometimes unable to handle, which indicates insufficient feature sets.

The described developed models are currently in use.

ACKNOWLEDGMENTS

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

REFERENCES

[1] K. Kenda, M. Škrjanc and A. Borštnik. *Modelling of the Complex Data Space.* Information, Intelligence, Systems, Applications, Corfu, July, 2015.

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

[3] 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.

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