Energy Efficiency
DOI 10.1007/s12053-015-9361-3

ORIGINAL ARTICLE

Using the Big Data generated by the Smart Home to improve energy efficiency management

María Rodríguez Fernández · Adolfo Cortés García · Ignacio González Alonso · Eduardo Zalama Casanova

Received: 3 March 2014 / Accepted: 25 May 2015

Abstract A Smart Home is able to generate energy-related values such as electricity consumption, temperature, or luminosity without higher infrastructure requirements. The main aim of this research is to extract information from that raw data that could contribute to improving the energy efficiency management. This paper presents a system which, using different Machine Learning approaches to learn about the users' consumption habits, is able to generate collaborative recommendations and consumption predictions that help the user to consume better, which will in turn improve the demand curve. Moreover, from consumption values, the system learns to identify devices, enabling the demand to be anticipated. Taking into account the fact that the amount of energy data is increasing in real-time, the use of Big Data techniques will be the key to handling all the operations and one of the more innovative features of the system.

M. Rodríguez Fernández (✉) · E. Zalama Casanova
University of Valladolid, Valladolid, Spain
e-mail: maria.rodriguez.fernandez@gmail.com

E. Zalama Casanova
e-mail: ezalama@eii.uva.es

A. Cortés García
Ingeniería de Integración Avanzadas (Ingenia) S.A, Málaga, Spain
e-mail: adolfo@ingenia.es

I. González Alonso
University of Oviedo, Oviedo, Asturias, Spain
e-mail: gonzalezaloignacio@uniovi.es

Keywords Smart meter · Sensor network · Power consumption · Machine learning · Energy efficiency · Big data

Introduction

The objective of increasing energy efficiency by 20 % was set in the European Union 2020 Strategy as a key factor to achieving long-term energy and climate goals (da Graça Carvalho 2012). Although substantial steps have been taken towards this objective, the European Commission estimated in 2009 that only half of the 20 % objective would be achieved if that trend continued and, therefore, a new energy efficiency [lan was developed in 2011. In this plan, the greatest energy-saving potential lies in buildings (nearly 40 % of the final energy consumption), with such policies as the creation of utilities to enable customers to cut their energy consumption (European Commission 2011).

In the current year, 2015, the Horizon 2020 program is being put into practice as the financial instrument implemented by the Innovation Union, a European 2020 flagship initiative aimed at securing Europe's global competitiveness. In this context, the concept of the Smart Grid is viewed as a key block and is represented in the main road maps throughout Europe (Massoud Amin and Wollenberg 2005).

The core element of Smart Grid is the active participation on the demand side, and this involves two main activities, load shifting and energy conservation programs (Palensky and Dietrich 2011). The first option

Published online: 02 June 2015

transfers customer load during periods of high demand to off-peak periods, flattening the load curve, while the second approach encourages customers to reduce their consumption, by such methods as reducing the air conditioning thermostat a few degrees.

Smart Metering (Stromback et al. 2011) is considered to be one of the most cost-effective methods for increasing end-consumer involvement and engagement. This method is based on an intelligent measuring device capable of reporting information about the power consumed. The said information can be managed by the final user to monitor and control the devices in their home, i.e., their own costs and expenses from any device connected to the Internet. If the management is optimum, the final bill is considerably reduced (Venables 2007). Another key enabling technology is the Sensor Network, made up of measuring devices that are distributed and embedded within the environment (Sheth et al. 2008), collecting such information as the temperature or humidity.

In the past, several attempts have been made to improve energy efficiency through the use of Smart Metering (Christine Easterfield 2013), and it is a fact that this type of infrastructure is becoming more widespread, although most of the information obtained from it is not being fully exploited (Jahn et al. 2010). The main aim of this research is to extract information from that raw data that can contribute to improving the energy efficiency management.

In that context, an architecture proposal able to reuse such data to give feedback, which is a possible proven energy saver (Fischer 2008), is presented in this research. The concept of Intelligent Infrastructure is applied—opening up the idea of Smart Grid (Gershenfeld et al. 2010)—and combined with the use of Machine Learning (ML) techniques, which permit learning to be done automatically from the data generated by the home devices, thus generating useful information to improve energy efficiency.

A prediction and recommender system can learn the consumption patterns of a home and thus contribute to the efficient use of energy. Such knowledge includes the technical aspects of behavior and habits of life, so a user can predict the consumption and adapt their activities to achieve more economies (for instance, considering the times with the best rate) and more environmentally responsible habits (Case 2012).

As well as saving energy, the system can help to detect fraud and abuse through consumption behavioral pattern analysis. For example, in communal areas, applying patterns to relate how much, when and how it is consumed, an improper use of the facilities could be detected. Furthermore, by modifying the consumption of final users, it is possible to flatten the demand curve, so the distribution network is optimized as a result of reducing consumption peaks. The knowledge of the behavior pattern of each house, every hour of the week, allows a better distribution of energy to meet demand. This could be generalized to any intelligent service development where low frequency data sampling may be necessary.

Concerning the possible growth of data, the traditional computing technologies have some limitations in terms of the capacity to store and process data, above which specific supercomputers are required, with a very high-associated cost. Big Data technology (Marz and Warren 2013) is able to approach the capabilities of supercomputers using conventional hardware, making it possible to apply the technology to fields in which it was unprofitable before. The use in real-time of Big Data and analytical data techniques applied in this context is the most innovative characteristic of this contribution.

The rest of the paper is organized as follows: firstly, in State of the art, an overview of the most relevant research that has made progress in improving energy efficiency or that has applied similar ideas and successful technologies to other fields. Secondly, a specific system applying the mentioned ideas will be described in Proposed system, with data about the initial results obtained. Finally, the document ends with the conclusions.

State of the art

A Smart Home can be defined as a complex system that integrates technology and services to enhance power efficiency and improve the quality of life (Robles and Kim 2010). It is necessary to take into account the fact that a growing number of consumer items (smart appliances, service robots, or electric vehicles) and different energy production sources (for instance, renewable energies like photovoltaic solar energy) are part of that home. Moreover, it is influenced by many sources of uncertainty, such as the outdoor temperature or the behavior of its occupants (Venkatesh 2008). In fact, user behavior is a key factor to explain households' energy consumption (Gram-Hanssen 2013). There are previous studies (Hargreaves et al. 2010) that show how feedback can change consumption behavior.

The design and development of energy management systems for homes would require the capability of predicting such parameters as the temperature or the energy demand to teach the user how to consume. Previous studies show that predictive models based on neural networks (González Lanza and Zamarreño Cosme 2002; González and Zamarreno 2005) and support vector machines (SVM) (Zhao and Magoulès 2012) are able to predict both the temperature evolution in a building and the consumption of their devices, allowing the preparation of corrective policies to improve the homes' energy efficiency. Regarding the requirements of energy demand, several research works have been carried out in the field of the identification of devices from power consumption using different algorithms of machine learning classification, which would anticipate the energy needs. However, the obtained results are not accurate enough (Berges et al. 2009; Murata and Onoda 2002). Similar classification and pattern characterization problems have been solved in other areas of knowledge, such as Computer Vision (Chechik et al. 2010) or the detection of malicious web sites (Ma et al. 2009), and the basis can be applied to this field of study.

Furthermore, taking into account the fact that the environment under study is repeated in all the Smart Homes and starting from the idea that two users/buildings with a similar energy profile could be interested in the same saving measures, a collaborative recommender system can be used as a complementary energy-saving tool. Specifically, one of the most used techniques for recommendations is collaborative filtering (CF), which filters items through the opinions of other people. It is based on the idea that if the advisors have similar preferences to the user, he/she is much more likely to take their opinions into account (Su and Khoshgoftaar 2009). This kind of tool is used successfully in other fields, such as the best known commercial online companies—e.g., Amazon.com (Linden et al. 2003). However, to the best of our knowledge, it has not yet been applied to the energy field. The fundamental assumption is that if users *X* and *Y* rate *n* recommendations similarly or have similar consumption behaviors (doing the washing at night, air conditioning during August, etc.), hence, they will rate or act on other recommendations similarly.

Regarding the state of the technologies for the realization of the ML process, Weka (Frank et al. 2010) is a matured software that has the advantage of incorporating a large number of training algorithms. It can be used for exploring representative data portions and testing the suitability of different learning methods easily. But for larger amounts of data, as in the case of the environment presented in this research, the tool did not offer a suitable response time, so specific tools for Big Data are needed. The Mahout Apache Project (Owen et al. 2011) and Jubatus Framework (Jubatus 2011) were selected for doing the batch and online learning, respectively. On the one hand, Mahout aims to produce a free implementation of a package that includes the main ML algorithms. The project is very active, but there are still some algorithms to be included, especially for time series classification. Its main advantage over other standalone implementations is the scalability offered when running on Hadoop (Lam 2010). On the other hand, the Jubatus online learning framework (Jubatus WebSite 2011) is a tool which maintains the scalability characteristics of Mahout and, in addition, allows a real-time response to be obtained. It incorporates some weighted algorithms which are the most appropriate for time series classifications.

Proposed system

The proposed system architecture is composed of four main sub-modules:

- The *Data Collection* element corresponds to the Smart Home Energy (SHE) project (SHE Consortium 2012), which obtains the measurements directly from the home.
- The collected values are processed by the *Data Storage* module, which makes them available for the rest of the modules.
- The *Machine Learning* module includes all the algorithms that can be executed massively for each home or user, many times a day, to learn, classify devices, generate recommendations, or predict the consumption using a large set of data for each execution.
- Finally, the *Data Visualization* module is the closest to the user and includes data publishing, allowing the interface to query, receive, and show the data.

The mentioned components will be described in depth in the following subsections.

Data collection

The Smart Home Energy environment, whose graphical description is shown in Fig. 1, is capable of obtaining energy-related information from the Digital Home, such as the power consumed by the electrical devices, the indoor temperature, or the luminosity in a specific zone.

Data acquisition is done by some hardware components—the Smart Meter Network and the Sensor Network—and a software adapter (SHE Adapter) that makes use of the Web Service for Device (WS4D) technology (Web Services for Devices (WS4D) Website (2012)) for making the measurement data available.

Before sending them to the Cloud, the SHE Adapter encapsulates the measurements obtained from the home into an object. Each measurement has an address, a location, and a time. Besides, for consumption values, the smart meter identification, the associated appliance, and the measured consumption value are also provided. In the case of the sensors, only the sensor code and its measure are needed.

Data storage

The collected energy-related data from the SHE environment is centrally stored by means of Cloud technology, which does not require a large infrastructure at home and, moreover, provides facilities to manage and maintain the integrity, security, and availability of data (Rhoton and Haukioja 2011).

The standardized energy measurement rate is 15 min (Franks 2012), due to the limitations of older technologies. If the rate increases, which is possible, thanks to Big Data, more detailed information would be offered, improving our understanding of what happens at home. Decreasing it, however, would represent a huge increase in the information storage and processing capacity, but further possibilities of extracting value from the information would be added. Furthermore, the incorporation of consumers to these measurement techniques would be a remarkable increase in communications, storage, and processing requirements.

Specifically, in the system, each home submits consumption information every minute, which means 525,600 samples per year. As other environmental parameters such as temperature or humidity are also measured, the number of samples per year in the system rises to more than three million. Each sample causes the transmission and storage of 1 kbps, so each home generates 3,229,286,400 bytes of information per year. Considering a population of ten million homes, a system able to process, receive, and store 32 PB of information per year is needed. Additionally, it would be necessary to support the execution of the operations to be performed with all this information, which could greatly increase the needs, and also to give support to remote user access.

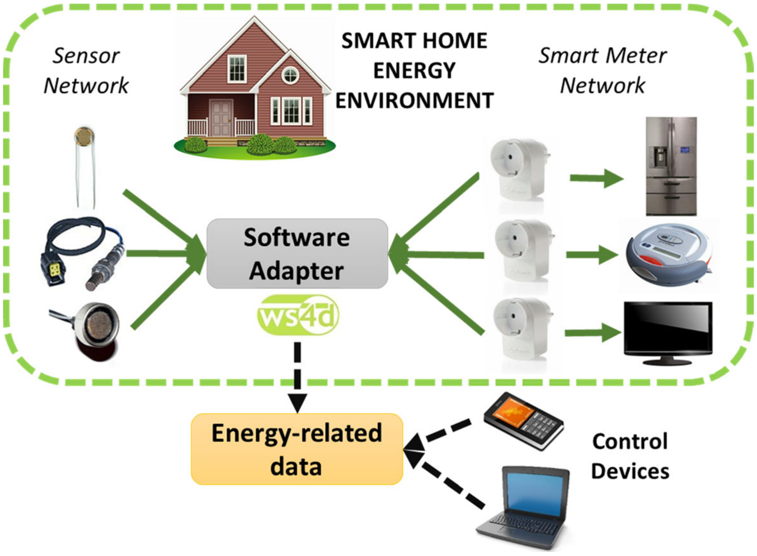


Fig. 1 Smart home energy environment

The storage needs are covered through a cost-linear investment technology based on Big Data. A distributed file system capable of storing up to the order of petabytes of information is the key to the storage management. The system self-manages the integrity of data through replication, without requiring backups or RAID disk enclosures. HBase technology is used to achieve real-time random access to databases consisting of large tables (billions of rows and millions of columns) through a Hive motor (Vora 2011).

Application of Machine Learning techniques to the collected energy-related data

Machine Learning is a technology for mining knowledge from data and then applying it to the new data. A common practice in machine learning to evaluate an algorithm is to split the data into two sets, the training set on which we learn data properties and the testing set on which we test these properties. Depending on the learning problem, the learning can be supervised (inputs and desired outputs are known) or unsupervised (unknown labels), and the technique is slightly different, so the learning details will be given in each subsection.

Appliance recognition module

This module's objective is to identify which device has the highest probability of generating a specific unlabeled consumption record. For this purpose, this module trains a classifier using supervised Machine Learning techniques.

Taking into account the fact that power consumption data is an infinite time series, the online learning approach was considered the most suitable option, offering simple, fast, and less-memory demanding solutions, avoiding re-training when adding new data since the model is generated incrementally.

The solution—graphically described in Fig. 2—follows a Jubatus client/server structure. The Classifier Tester processes the Big Data stream and composes the Training Datum (nomenclature used by Jubatus). Specifically, it is made up of the minimum and maximum values, the mean, sum, deviation, and variance. Finally, the Fast Fourier Transform (FFT) was used to introduce the frequency aspect.

The classification process involves two main operations that can be performed in parallel (50 % of the measured data is used for each one):

- *Training the model*. The *train* function of the Classifier Client receives the labeled record (list of tuple of datum and its label, following the Jubatus nomenclature) and returns the number of trained datum. In each of the iterations, the Classifier Server obtains and saves a new version of the model.
- *Classification*. The model is tested with unlabeled data. It receives the list of datum to classify and returns a list of estimated results, specifically, for each label, the probability of having generated the input record. The score, which represents the possibility of belonging to each class, is calculated as the inner product of the model coefficients and the feature vector. In this step, it is possible to adjust the

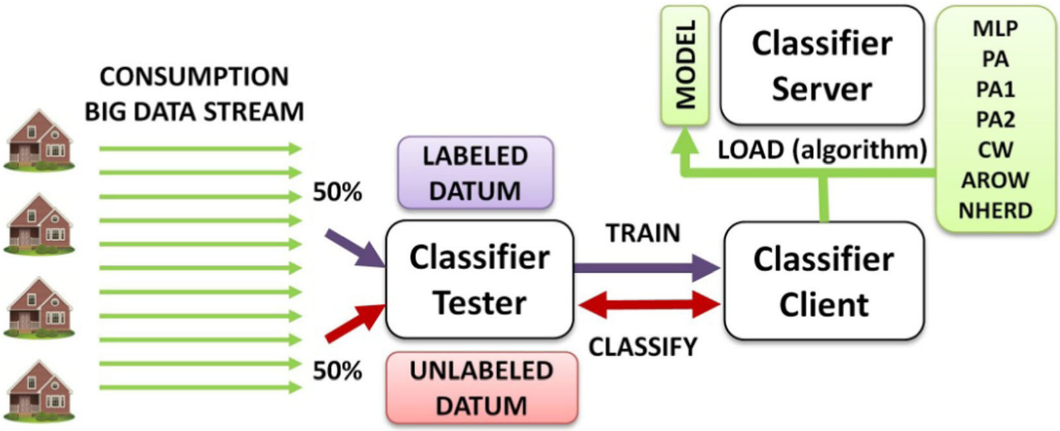


Fig. 2 Appliance identification process

input variables and configure the learning algorithm in order to improve the accuracy of the classifier.

The accuracy of the classifier is calculated as the number of times that an unlabeled record is classified in the correct class in relation to the total number of tests carried out. Both the first and second choice were taken into account, considering the weight of the second as half (Stamatatos and Widmer 2005).

The results shown in this work were obtained by training the Jubatus classifier with the consumption values of seven appliances situated in the Smart Home (CRT monitor, LCD monitor, heater, lamp, fridge, printer, and smart TV), sending measures at a 1-min rate (hence, the duration of each training iteration is 16 min). The experiment has tested all the training algorithms implemented in Jubatus. They can be classified in three families according to their basis:

- Perceptron (McDonald et al. 2010), the classical online learning algorithm performs a multi-class classification based on a set of weight vectors (one for each class), which are updated according to the prediction results, leading to the segmentation of the data space. Based on this, Passive Aggressive (PA) (Crammer et al. 2006) is offered by the tool in three different implementations (PA, PA1, and PA2), but it does not offer optimum results in multi-class classification.
- Confidence Weighted (CW) (Crammer et al. 2009a, b) is based on the notion of a parameter confidence measure, as an improvement over the aforementioned methods. Maintaining this idea, Adaptive Regularization of Weights (AROW) (Crammer et al. 2009a, b) also offers large margin training and the capacity to handle non-separable data. Normal Gaussian herding (NHERD) (Crammer and Lee 2010) is an attempt to improve learning when some noise is present.

The evolution learning rate of the classifier for each algorithm can be seen in Fig. 3, where the accumulated accuracy in the first 15 iterations (4 h of training) is shown. The best performance (about 74 % recognition rate) was obtained using weighted algorithms, such as CW and AROW, which, moreover, are the fastest, presenting values above 60 % of success after 30 min of training.

Collaborative recommender

The function of this module is to suggest actions to the users that other similar users have done in their homes, helping them to reduce their energy consumption.

To this end, it will be necessary to define a method for calculating the similarity between two users, i.e., the "distance" between them. When two users are "close" enough, they are considered as belonging to the same

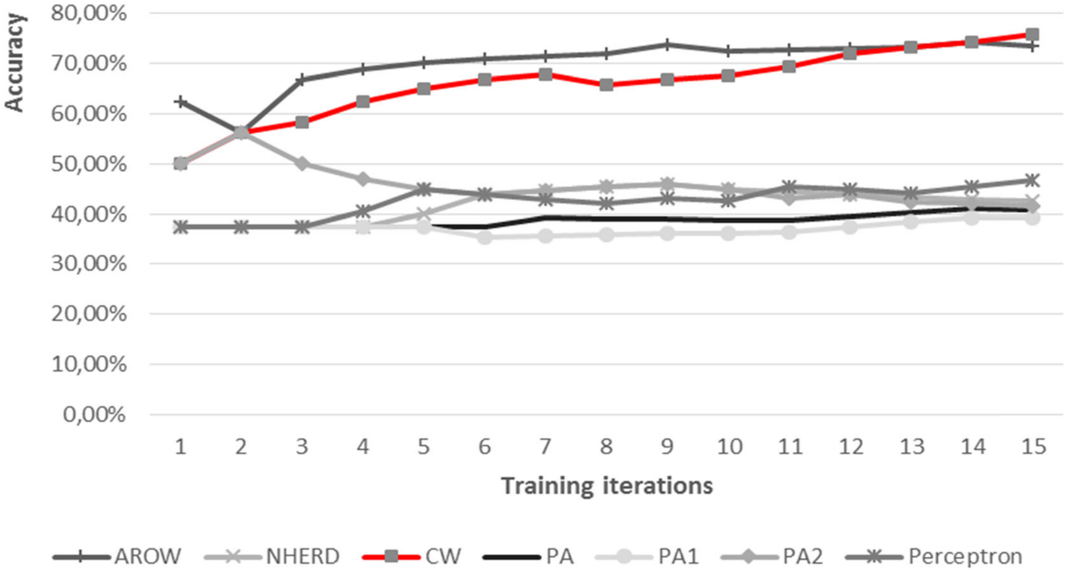


Fig. 3 Evolution of the accumulated accuracy of the classifier for each tested training algorithm

"neighborhood"—virtually speaking—and therefore similar recommendations are offered to them.

The Pearson correlation coefficient and the Euclidean distance are two valid values, which are based on the degree of acceptance of the actions that the users chose. In the specific solution proposed in this work, the users can agree (value 0) or not (value 1) with the completion of an action. In this context, the most appropriate algorithms are those of Tanimoto (Cechinel et al. 2013). The Tanimoto similarity coefficient can be expressed as (1), where *A* and *B* are the number of actions carried out by the two users whose similarity is calculated, respectively, and *C* corresponds to the number of actions common to both users.

$$T(A,\ B) = \frac{C}{A + B - C} \tag{1}$$

From the formula, the distances between all the users are calculated and stored in a matrix. When a recommendation is required to be given to the user, the algorithm returns the actions that most similar users have done and that the user has not yet carried out.

As for implementation, specific tools such as Mahout already incorporate these algorithms. The basic steps are to build a recommendation engine, and from there, generate recommendations. A sample code can be seen in Fig. 4.

Predictions

The system generates a weekly consumption prediction for each user. This massive information can be handled because the algorithms are implemented using Big Data technologies, which enable the code to be parallelized and the calculations to be performed for N users in a distributed way (cluster of machines), with a linear function of associated costs.

To make this kind of prediction, which can be seen in Fig. 5, specific and overall consumption factors are taken into account, as well as other relevant external factors, such as holiday schedules, prices, or meteorology. The graph allows the user to zoom over it and to compare the consumption (in green) with the prediction (in blue). The temperature (°C) is also shown (in red) as it is a representative input to guide the learning process, since the air conditioning consumption is a relevant parameter in business buildings and offices.

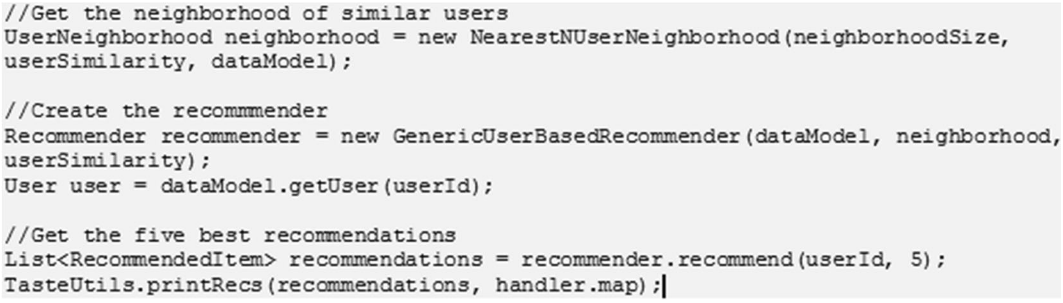
Regarding the methodology, the system tries to learn the behavior under certain climatic conditions, also considering social or behavioral aspects. In the first iteration, the variations in the prediction were analyzed and different patterns of day were found (holidays, holidays with commercial time schedule, holidays with full business hours, etc.). This point was solved by typifying the days and creating different units of knowledge for each of them.

To consider the seasonal variation, the system first determines whether the type of day corresponds to a day in winter, spring, summer, or fall, depending on the weather, and applying the acquired knowledge of the season.

Therefore, the system learns and predicts a differentiated "unit of knowledge" corresponding to the time range (morning, noon, afternoon, evening, night), the range type (spring morning, noon summer, etc.), the type of day (weekday morning spring, festive, festive opening noon), and the day except holidays (Monday spring morning business, etc.), obtaining 140 units of knowledge.

The system could determine which ML algorithm is applied in each unit of knowledge, which is different from one user to the next and even some seasons of the year, schedules, and other ranges.

The system learns in each case, i.e., how the user behaves on Monday mornings in spring, on Sunday



```
//Get the neighborhood of similar users
UserNeighborhood neighborhood = new NearestNUserNeighborhood(neighborhoodSize,
userSimilarity, dataModel);

//Create the recommmender
Recommender recommender = new GenericUserBasedRecommender(dataModel, neighborhood,
userSimilarity);
User user = dataModel.getUser(userId);

//Get the five best recommendations
List<RecommendedItem> recommendations = recommender.recommend(userId, 5);
TasteUtils.printRecs(recommendations, handler.map);
```

Fig. 4 Sample code for generating collaborative recommendations

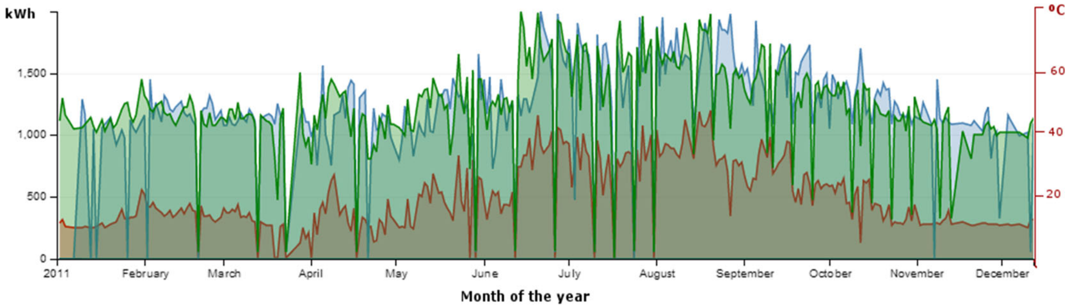


Fig. 5 Expandable graphic prediction generated by the systems, in comparison with the real measure

afternoons in fall, etc., and depending on the day to predict, the system selects the appropriate unit of knowledge, which relearns from the new measurement.

The strategy used by the model allows the way in which the system learns to be defined and adjusted, enhancing the use of relevant data and filtering the irrelevant data.

The model has been trained and tested over 2 years—one for training and one for testing—using the information (temperature, humidity, and consumption) generated by two buildings located in different climatic zones, specifically, Atlantic (a seaside city of northern Spain, Vigo) and Continental (Madrid, located in the center of the country) climate zones. The model is able to learn from the data how energy consumption is correlated with humidity and temperature measurements and then get the prediction for 1 week ahead. The value used for testing is actually the corresponding measured value for the second year. The effectiveness of the prediction system has been evaluated by statistical analysis. The interpretation of these statistics permits the error in the prediction to be calculated and the way in which that error was distributed to be evaluated and thereby adjust the learning model.

From the results listed in Table 1, some conclusions can be drawn. The MAD value offers an insight to the error that can occur in a prediction, and it is similar in both climatic zones. Besides that, MAPE, defined as the probability (between 0 and 1) of the average prediction errors, presents a better value in the oceanic climatic zone. The PA can be calculated from MAPE to compare different datasets. This technique shows a better value in the oceanic climatic zone. In the context of our system, we can predict with that confidence how much and in which way a user has consumed, taking the recorded consumption of last year. Considering the RMSD values, the errors are well spread over time if the value is low or very concentrated at certain times if the value is high. Having concentrated errors can be interpreted as more convenient for adjusting to anomalies. The oceanic zone shows a greater concentration of error than the continental zone, according to this technique. Finally, the SD is considered as an indicator of the error dispersion regarding the average error. Since the error is concentrated at certain points, the algorithm has a more stable performance regarding the error committed in each prediction than if the error was uniformly distributed because it would mean that wide variations exist between different predictions. The continental zone offers a better value in that case.

Data mining and insight

The information is presented to the user using highly expressive visual graphics, thanks to Data Mining and Insight technologies, which permit the user to view a lot of varied information at a glance, thus avoiding having

Table 1 System efficiency evaluation results by means of statistics

Statistic	Atlantic climate zone	Continental climate zone
Median absolute deviation (MAD)	237.46 kWh	207.59 kWh
Mean absolute percentage error (MAPE)	0.099	0.194
Prediction accuracy (PA)	90.09 %	80.59 %
Root mean square deviation (RMSD)	333.03 kWh	281.17 kWh
Standard deviation (SD)	233.49 kWh	189.63 kWh

many disjoint element graphics. In addition, inspection capabilities are provided, allowing the user to select sections, filtering the remaining information in order to draw conclusions.

A relevant graph is the “Heat Map” of consumption, where the consumption is represented by a disk comprising seven concentric circles (one for each day of the week) and 24 segments (one for each hour of the day), wherein the color corresponds to the intensity of consumption (to higher consumption, greater intensity of blue). This graph helps the user to know his/her consumption pattern, showing the differences in consumption between weekdays, which can help to improve the associated habits. Moreover, it facilitates the detection of anomalies that may be caused by device failures, which cause higher consumption. For example, Fig. 6 shows a higher consumption from 0700 to 2100 h on weekdays. Besides, two consumption anomalies can be distinguished, a fall on Saturday between 0200 and 0300 h and a rise on Monday between 0000 and 0100 h.

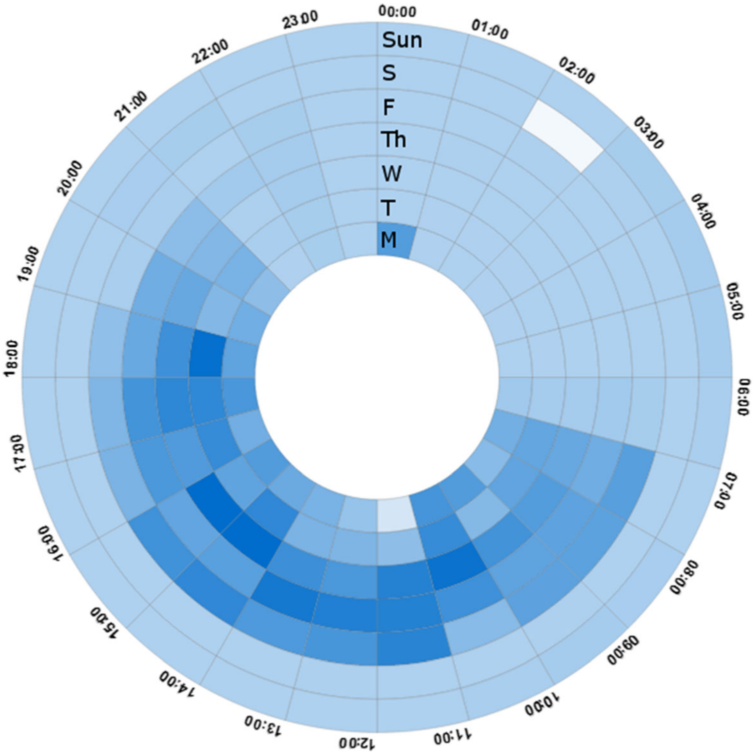
Relations between the modules

The elements described in the current section can be seen graphically in the SysML Block Diagram of Fig. 7 (dotted rectangles), which includes both hardware and software elements.

At the top of the hierarchy, the *SHE Adapter* collects the values such as the consumption from the home. The obtained data is then processed by the *Data Storage* module, which follows the concepts of Lambda Architecture (Fan and Bifet 2013), decomposing the problem into three layers:

- The *Acquisition block* collects the data from the SHE Adapter, to make it available for the Batch and Real-time blocks.
- The *Real-time block* is needed to provide a real-time monitoring and control services to users through a Graphical User Interface. This layer needs to aggregate the data, using typical functions, such as average and summarization for each user, group of devices, and time intervals, in a near real-time continuous computing.
- The *Batch block* includes several components to store and process the data, applying Data Mining and Machine Learning algorithms to acquire a customized knowledge pool for the home energy consumption.

Fig. 6 Heat Map of air conditioning consumption



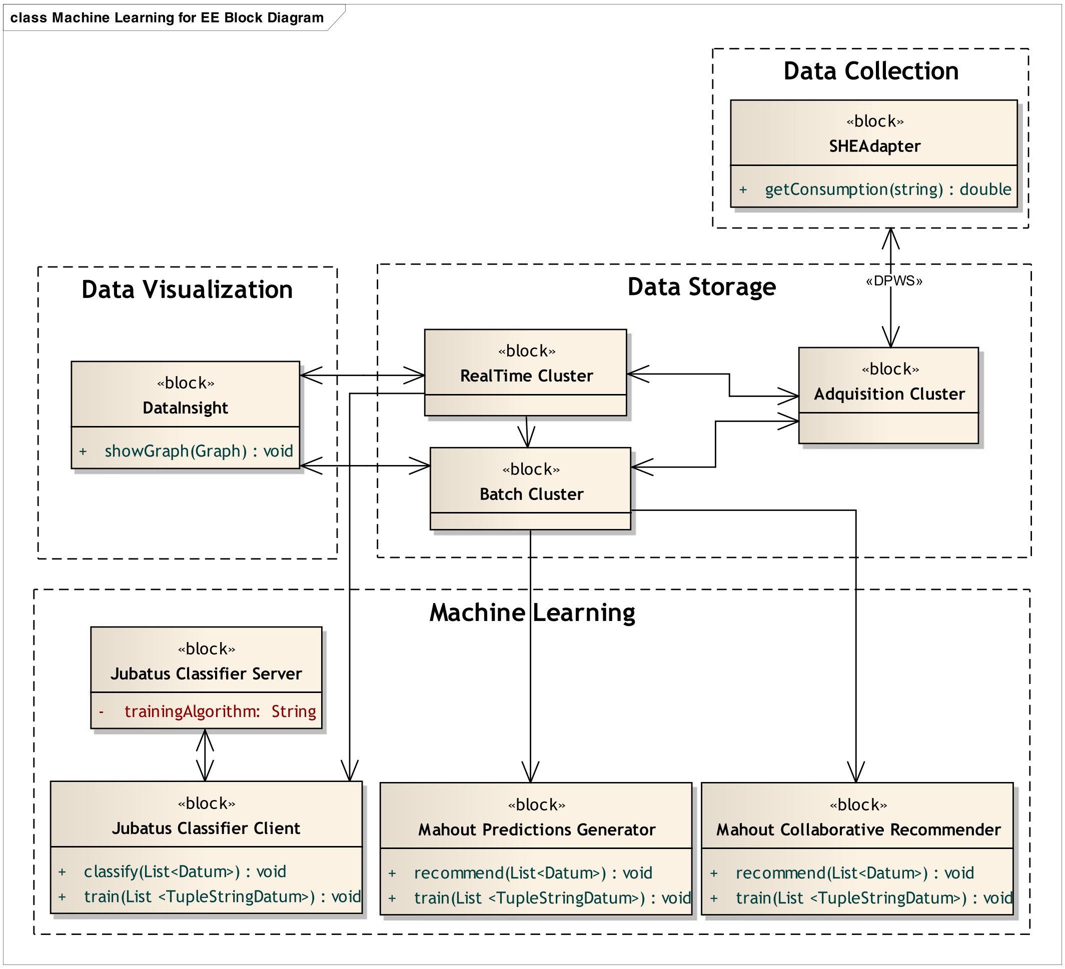

class Machine Learning for EE Block Diagram
Data Collection
«block»
SHEAdapter
+ getConsumption(string) : double
«DPWS»
Data Visualization
«block»
DataInsight
+ showGraph(Graph) : void
Data Storage
«block»
RealTime Cluster
«block»
Adquisition Cluster
«block»
Batch Cluster
Machine Learning
«block»
Jubatus Classifier Server
- trainingAlgorithm: String
«block»
Jubatus Classifier Client
+ classify(List<Datum>) : void
+ train(List <TupleStringDatum>) : void
«block»
Mahout Predictions Generator
+ recommend(List<Datum>) : void
+ train(List <TupleStringDatum>) : void
«block»
Mahout Collaborative Recommender
+ recommend(List<Datum>) : void
+ train(List <TupleStringDatum>) : void

Fig. 7 Proposed system SysML block diagram

The stored data is then available to be used both by the *Data Insight* block to generate interactive graphs and *Mahout* and *Jubatus* instances, which will execute the learning over them, as was described earlier in the section.

Conclusions

The knowledge of household energy is a key factor in achieving the efficient use of resources. The widespread use of Smart Meters and Sensor Networks at residential level facilitates the obtaining of data, but due to their variety and size, it cannot be directly used to make conclusions that help to improve the energy efficiency.

The architecture of a four-module system based on Machine Learning techniques, combined with Big Data technology, has been presented in this work. Big Data allows large volumes of varied data to be managed and offers support for ML algorithms, Data Mining visual tools, near to real-time monitoring, and other information analysis and processing possibilities that fit perfectly with the requirements.

The Data Collection module is based on the data generated by the Smart Home Energy project, so the solution does not require investment in infrastructure. Moreover, it could be applied to any other similar smart environments.

Cloud technology offers an elastic and resilient solution without requiring a high-capacity storage infrastructure at the household level. Besides, the layered design allows both a batch and a continuous real-time processing of the measurements to be done, working with a large set of data taken over time from a large set of homes and historical database (Data Storage module).

The Machine Learning module is composed of three elements. First, a supervised classifier is trained to recognize each device from the consumption data, with the aim of being able to anticipate its consumption demand. By using some weighted algorithms, recognition rates above 74 % have been obtained, a value which improves with time, since the learning is done online. A potential improvement could be the use of clustering techniques in a previous phase in order to find out which category the device belongs to and therefore reduce the number of candidates that the classifier would need to screen it with. Furthermore, by applying the concept of user energy profile, a collaborative recommender processes the user actions in order to make energy-saving suggestions for similar users. It is also possible to extract consumption patterns and thus allow predictions to be made to anticipate and adapt to other cheaper options. The efficiency of this module was evaluated in buildings located in two different climatic zones, and an accuracy of 90 % has been achieved in the Atlantic Zone.

A useful complementary tool—belonging to the Data Mining and Insight module—, is the incorporation of interactive and customizable graphs to show the information to the user, who is able to manage the energy consumption and consequently improve the energy efficiency.

Summarizing, a complete infrastructure to improve the energy efficiency from the data generated by a smart environment has been proposed. The main advantages of the solution are that it is open, distributed, and scalable. The application of Big Data technology allows the information to be analyzed in more detail than with traditional technology, and the application of it to the energy sector is an innovative idea.

Acknowledgments This work is funded by the Ministry of Economy and Competitiveness of Spain (IPT-2011-1237-920000) and FEDER funds. We are also very grateful to all the members of the SHE consortium (Ingenia, Satec, Ingho, Tecopysa, Cotesa, IAT, and the University of Oviedo). The authors would like to thank the University of Oviedo for hosting this research and especially the staff of the Infobotica Research Group (University of Oviedo) for all their valuable comments.

References

Berges, M., Goldman, E., Matthews, H. S., Soibelman, L. (2009). Learning systems for electric consumption of buildings. In *ASCI international workshop on computing in civil engineering*. http://ascelibrary.org/doi/abs/10.1061/41052(346)1. Accessed 31 October 2014.

Case, R. (2012). Saving electrical energy in commercial buildings. https://www.uwspace.uwaterloo.ca/handle/10012/6885. Accessed 14 April 2015.

Cechinel, C., Sicilia, M.-Á., Sánchez-Alonso, S., & García-Barriocanal, E. (2013). Evaluating collaborative filtering recommendations inside large learning object repositories. *Information Processing and Management, 49*(1), 34–50. doi:10.1016/j.ipm.2012.07.004.

Chechik, G., Sharma, V., Shalit, U., & Bengio, S. (2010). Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research, 11*, 1109–1135. Accessed 31 October 2014.

Crammer, K., & Lee, D. D. (2010). Learning via gaussian herding. *Pre-proceeding of NIPS*. http://webee.technion.ac.il/Sites/People/koby/publications/gaussian_mob_nips10.pdf. Accessed 6 June 2013.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research, 7*, 551–585. Accessed 31 October 2014.

Crammer, K., Dredze, M., Kulesza, A. (2009). Multi-class confidence weighted algorithms. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 2-Volume 2* (pp. 496–504). http://dl.acm.org/citation.cfm?id=1699577. Accessed 6 June 2013.

Crammer, K., Kulesza, A., Dredze, M. (2009). Adaptive regularization of weight vectors. *Advances in Neural Information Processing Systems*, *22*, 414–422. http://www.cis.upenn.edu/~kulesza/pubs/arow_nips09.pdf. Accessed 6 June 2013.

Da Graça Carvalho, M. (2012). EU energy and climate change strategy. *Energy, 40*(1), 19–22. doi:10.1016/j.energy.2012.01.012.

Easterfield, C. (2013). The customer's impact in smart metering. http://www.european-utility-week.com/Pages/Detail/6190. Accessed 1 July 2013.

European Commission. (2011). *Energy efficiency plan 2011*. Brussels.

Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter, 14*(2), 1–5. Accessed 25 October 2014.

Fischer, C. (2008). Feedback on household electricity consumption: a tool for saving energy? *Energy Efficiency, 1*(1), 79–104. doi:10.1007/s12053-008-9009-7.

Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., Trigg, L. (2010). Weka—a machine learning workbench for data mining. In *Data mining and knowledge discovery handbook* (pp. 1269–1277). Springer. http://link.springer.com/chapter/10.1007/978-0-387-09823-4_66. Accessed 5 June 2013.

Franks, B. (2012). *Taming the big data tidal wave: finding opportunities in huge data streams with advanced analytics* (Vol. 56). Wiley. com. http://books.google.es/books?hl=es&lr=&id=-oPQrEQzTAsC&oi=fnd&pg=PR13&dq=Taming+The+Big+Data+Tidal+Wave:+Finding+Opportunities+in+Huge+Data+Streams+with+advanced+analytics&ots=hVfXqIpGAI&sig=7st_s1FnJ2grMEipg6FVb7CO124. Accessed 5 November 2013.

Gershenfeld, N., Samouhos, S., & Nordman, B. (2010). Intelligent infrastructure for energy efficiency. *Science, 327*(5969), 1086–1088. doi:10.1126/science.1174082.

González Lanza, P. A., & Zamarreño Cosme, J. M. (2002). A short-term temperature forecaster based on a state space neural network. *Engineering Applications of Artificial Intelligence*, *15*(5), 459–464. http://www.sciencedirect.com/science/article/pii/S0952197602000891. Accessed 6 February 2013.

González, P. A., & Zamarreno, J. M. (2005). Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. *Energy and Buildings*, *37*(6), 595–601. http://www.sciencedirect.com/science/article/pii/S0378778804003032. Accessed 11 July 2013.

Gram-Hanssen, K. (2013). Efficient technologies or user behaviour, which is the more important when reducing households' energy consumption? *Energy Efficiency, 6*(3), 447–457. doi:10.1007/s12053-012-9184-4.

Hargreaves, T., Nye, M., & Burgess, J. (2010). Making energy visible: a qualitative field study of how householders interact with feedback from smart energy monitors. *Energy Policy, 38*(10), 6111–6119. Accessed 17 October 2014.

Jahn, M., Jentsch, M., Prause, C. R., Pramudianto, F., Al-Akkad, A., Reiners, R. (2010). The energy aware smart home. In *2010 5th International conference on future information technology (FutureTech)* (pp. 1–8). Presented at the 2010 5th International Conference on Future Information Technology (FutureTech). doi:10.1109/FUTURETECH.2010.5482712.

Jubatus. (2011). Jubatus: real-time and highly-scalable machine learning platform. *Hadoop summit 2013 North America: Community choice now open!*. http://hadoopsummit2013.uservoice.com/forums/196822-future-of-apache-hadoop/suggestions/3714873-jubatus-real-time-and-highly-scalable-machine-lea. Accessed 5 June 2013.

Jubatus WebSite. (2011). Jubatus : distributed online machine learning framework—Jubatus. http://jubat.us/en/. Accessed 19 June 2013.

Lam, C. (2010). *Hadoop in action* (1st ed.). Manning Publications.

Linden, G., Smith, B., York, J. (2003). Amazon. com recommendations: item-to-item collaborative filtering. *Internet Computing, IEEE*, *7*(1), 76–80. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1167344. Accessed 3 July 2013.

Ma, J., Saul, L. K., Savage, S., Voelker, G. M. (2009). Identifying suspicious URLs: an application of large-scale online learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 681–688). http://dl.acm.org/citation.cfm?id=1553462. Accessed 6 June 2013.

Marz, N., & Warren, J. (2013). *Big data: Principles and best practices of scalable realtime data systems*. Manning Publications.

Massoud Amin, S., & Wollenberg, B. F. (2005). Toward a smart grid: power delivery for the 21st century. *Power and Energy Magazine, IEEE*, *3*(5), 34–41. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1507024. Accessed 1 July 2013.

McDonald, R., Hall, K., Mann, G. (2010). Distributed training strategies for the structured perceptron. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 456–464). http://dl.acm.org/citation.cfm?id=1858068. Accessed 6 June 2013.

Murata, H., & Onoda, T. (2002). Estimation of power consumption for household electric appliances. In *Neural information processing, 2002. ICONIP'02. Proceedings of the 9th international conference on* (Vol. 5, pp. 2299–2303). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1201903. Accessed 19 December 2012.

Owen, S., Anil, R., Dunning, T., Friedman, E. (2011). *Mahout in action* (Pap/Psc.). Manning Publications.

Palensky, P., & Dietrich, D. (2011). Demand side management: demand response, intelligent energy systems, and smart loads. *IEEE Transactions on Industrial Informatics, 7*(3), 381–388. doi:10.1109/TII.2011.2158841.

Rhoton, J., & Haukioja, R. (2011). *Cloud computing architected: solution design handbook.* Recursive Press.

Robles, R. J., & Kim, T. (2010). Applications, systems and methods in smart home technology: a review. *International Journal of Advanced Science and Technology, 15*, 37–47. Accessed 31 October 2014.

SHE Consortium. (2012). Smart home energy. http://156.35.46.38/she/. Accessed 8 January 2013.

Sheth, A., Henson, C., Sahoo, S. S. (2008). Semantic sensor web. *Internet Computing, IEEE*, *12*(4), 78–83. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4557983. Accessed 1 July 2013.

Stamatatos, E., & Widmer, G. (2005). Automatic identification of music performers with learning ensembles. *Artificial Intelligence, 165*(1), 37–56. Accessed 22 October 2014.

Stromback, J., Dromacque, C., Yassin, M. H., VaasaETT, G. E. T. T. (2011). The potential of smart meter enabled programs to increase energy and systems efficiency: a mass pilot comparison short name: empower demand. *Vaasa ETT.* http://www.bwrassociates.co.uk/vaasaett/wp-content/themes/blue-grace/images/Final_Empower.pdf. Accessed 1 July 2013.

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence, 2009*, 4. Accessed 17 October 2013.

Venables, M. (2007). Smart meters make smart consumers [Analysis]. *Engineering Technology, 2*(4), 23–23.

Venkatesh, A. (2008). Digital home technologies and transformation of households. *Information Systems Frontiers, 10*(4), 391–395. doi:10.1007/s10796-008-9097-0.

Vora, M. N. (2011). Hadoop-HBase for large-scale data. In *Computer science and network technology (ICCSNT), 2011 international conference on* (Vol. 1, pp. 601–605). IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6182030. Accessed 25 October 2014.

Web Services for Devices (WS4D) Website. (2012). http://ws4d.e-technik.uni-rostock.de/. Accessed 28 May 2013.

Zhao, H., & Magoulès, F. (2012). A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, *16*(6), 3586–3592. http://www.sciencedirect.com/science/article/pii/S1364032112001438. Accessed 6 February 2013.