Journal of Computing and Information Technology - CIT 18, 2010, 4, 341–347
doi:10.2498/cit.1001913

Using Machine Learning on Sensor Data

Alexandra Moraru[1], Marko Pesko[1,2], Maria Porcius[3], Carolina Fortuna[1] and Dunja Mladenic[1,3]

[1] Jožef Stefan Institute, Ljubljana, Slovenia
[2] MOBITEL Telecommunication Services Inc., Ljubljana, Slovenia
[3] J. Stefan International Postgraduate School, Ljubljana, Slovenia

Extracting useful information from raw sensor data requires specific methods and algorithms. We describe a vertical system integration of a sensor node and a toolkit of machine learning algorithms for predicting the number of persons located in a closed space. The dataset used as input for the learning algorithms is composed of automatically collected sensor data and additional manually introduced data. We analyze the dataset and evaluate the performance of two types of machine learning algorithms on this dataset: classification and regression. With our system settings, the experiments show that augmenting sensor data with proper information can improve prediction results and also the classification algorithm performed better.

Keywords: sensor node, data mining, machine learning, prediction

1. Introduction

The development of sensor networks, particularly in the last years, has extended their applicability in various domains, such as heritage preservation, environmental motoring and human activity recognition. Sensor based systems are known as being highly application dependent. This also includes the top layer, which should handle the data in an efficient and useful manner. Furthermore, the size of collected data is rapidly increasing with the number and scale of deployed sensor networks and specialized methods able to deal with such scale and still satisfy application requirements are needed.

In this paper, we show how we can apply machine learning (ML) algorithms on automatically gathered sensor data combined with manually collected data in order to predict different events. Our demonstration is based on the data collected from a sensor node deployed in our lab, which measured temperature, humidity, light and pressure over 15 days. These parameters are affected by human presence. In parallel, we manually collected data related to human presence and events in the lab. These two sets of data are then aligned and used for training ML algorithms which are then able to predict the number of people in the lab.

This work is focused on ML for analysis of sensor data as a part of a complete vertical system integration, spanning from hardware at the bottom level to data-driven ML algorithms at the topmost. To the best of our knowledge, this is the first sensor system with such a deep vertical integration. Moreover, we consider our system as an example of applying machine learning methods on sensor data, which can provide high-level guidelines for similar applications involving prediction from sensor data.

Besides the direct applicability of the system in predicting the number of people in closed spaces, target applications can also be in the area of museums, libraries and protected buildings where predicting the number of people in a hall is vital for preserving valuable heritage [1].

The rest of this paper is structured as follows. Section 2 gives system description, Section 3 provides interpretation and evaluation of the results, while Section 4 presents related work. In Section 5 we draw the conclusion and give some directions for future work.

2. The Vertical System

The components of the system are presented in Figure 1 and they are: the sensor node, a server collecting sensor data, a human component for introducing additional data, database with the additional data, data preprocessing tools and ML toolkit. For gathering the dataset, first we get raw data from sensors on the sensor node, and then we transmit these data to a machine for storage. In the next step, we use preprocessing methods to integrate automatically collected data with manually labeled data for training ML algorithms.

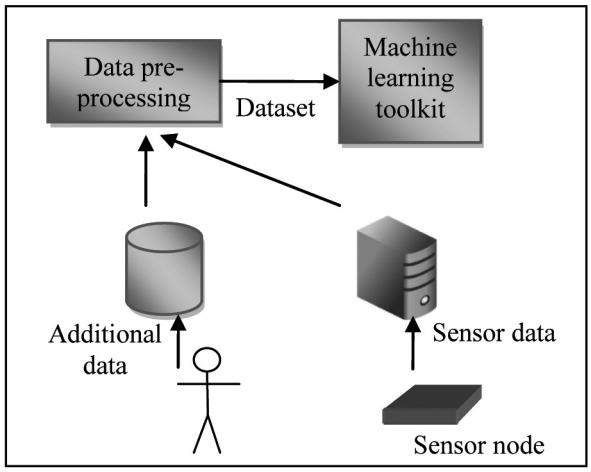

Figure 1. Vertical system integration.

2.1. Sensor node

The sensor node consists of sensors, power supply, LCD and ATmega128L microcontroller connected to the PC via RS232 to USB converter. We used Taos TSL2561 light-to-digital converter for measuring luminance, Sensirion SHT11 for temperature and relative humidity and VTI SCP1000 for absolute air pressure. The microcontroller gathers data from sensors at a sampling rate of 10 seconds, then packs data into a vector and finally sends it to a server for storage, via the serial port.

2.2. Data gathering

Sensor data are read from the serial port by an application (VS 2005 .NET application written in C#) which allows the user to set custom port, baud rate, target storage (database or simple text file) through a GUI. In our experiments, data were stored in a text file, each sample containing 4 numerical values (temperature, humidity, light and pressure) and a time stamp. Table 1 contains the statistics of sensor data: minimum and maximum values, and mean value calculated from all instances. From this, it can be concluded that the measurements are correct, with no extreme values.

	Min	**Max**	**Mean**
Temperature (°C)	16.94	29.93	25.93
Humidity (%)	14.43	47.9	31.12
Light (Lux)	0	64.71	26.30
Pressure (hPa)	995.7	1025.6	1011.9

Table 1. Sensor data statistics.

In addition, three more attributes have been manually entered in the separate table: the number of people in the lab, the number of computers running and the position of the window. Each of these attributes has attached a time stamp, whenever a change in their values occurred. In this process of collecting additional data, all the persons working in our lab were involved. The class attribute is represented by the number of people present at one time in the lab, and it can have one of the following values: 0, 1, 2, 3.

Figure 2 plots an example of the variation of sensor data in the lab during a working day. From aligning this data with the additional manually introduced data (see the sample in Table 2) we

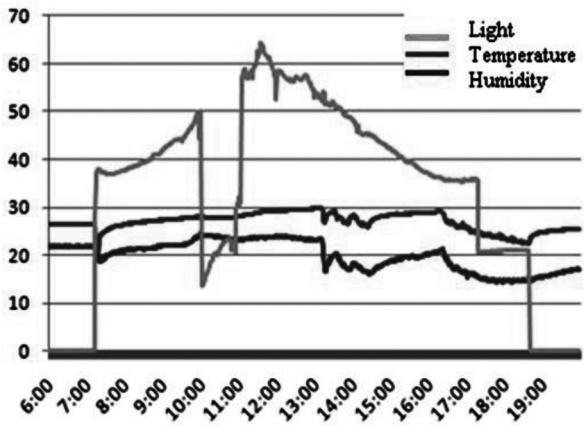

Figure 2. Sensor data variation.

Time	Persons	Window	Computers
7:15	1	open	3
10:00	0	closed	3
11:04	3	closed	3
13:11	3	open	3
13:55	3	closed	3
14:09	2	1/2 open	3
18:39	0	closed	2

Table 2. Manually collected data sample.

can observe that the humidity and temperature values are rising with the increase in the number of people in the lab. Also, the intensity of the light is higher when people are in the lab, mainly due to artificial illumination. For example, it can be observed that the light was rapidly increasing at 7:15 when a person entered the room and turned on the lights. Significant changes in the temperature and humidity trend lines appear when the window is opened.

2.3. Data processing and learning

The first step in data processing was the alignment of sensor data with the manually collected data, based on time stamps, resulting in an augmented dataset. The initial sampling rate for the sensor data was of 10 seconds, while for the manually collected data, the time stamp contained only the hour and minute of the entry. For all instances from the sensor data within the same minute we assigned the corresponding instance from the additional collected data.

In the second step we performed dataset reduction, first by choosing a sampling rate of 1 minute, since the manually collected data has the time stamp only in hour and minutes. However, some of the features from the sensor data (i.e. ambient light) present high variations during one minute, which cannot be correctly correlated with additional data. Namely, the moment when a person enters the office and turns on the light is sensed in no more than 10 seconds by the sensor device, while in the additional data, this is marked only in minutes. Moreover, the difference of the time stamps for the two sources of data may vary with a few minutes, because more people entered the additional data and was no time synchronization applied. We have also eliminated the data obtained during the night (between 8:00 PM and 6:00 AM), to avoid too many instances with 0 persons. Namely, the data obtained during the night had no persons and including all these data would make our dataset very unbalanced. Therefore, the resulting dataset contains some incorrect instances due to human errors and the impossibility of perfectly aligning sensor data with manually collected data. On the other hand, there are no missing values in the dataset.

The dataset used in the learning process contains a total of 16,578 measurements, each with 9 attributes: temperature, humidity, light, pressure, weekday (with two nominal values: working day and weekend), hour interval (integer values between 6 and 19), position of the window (with three nominal values: open, half open and closed), number of computers working (integer values between 1 and 4) and number of people in the lab (class attribute). The value distribution for the class attribute is: 44.17% instances with 0 persons, 25.27% instances with 1 person, 22.24% instances with 2 persons, 8.32% instances with 3 persons, respectively.

On this dataset we applied two learning methods: classification and regression. The first method is used for predicting categorical class labels, while the second method models continue-valued function for approximating the target variable (class attribute). We decided to test both methods since the target variable, in our settings, can be seen both as a variable with discrete values or as a numerical variable. The classification algorithms applied are decision tree and Bayesian network. The first one provides a good visual interpretation of the results and the second one has a better usage of all the attributes of the dataset. As a regression algorithm we chose the commonly used linear regression.

3. Interpretation and Evaluation of the Results

In order to evaluate the power of prediction of each algorithm, we conducted experiments on two cases; the simple case is when the dataset contains only sensor data attributes (temperature, humidity, light and pressure) and the class

Evaluation \ Algorithm		J48	BayesNet	Linear Regression
Simple Dataset	MAE	0.17	0.12	0.44
	RMSE	0.29	0.26	0.54
	ACC	73%	80%	—
Augmented Dataset	MAE	0.15	0.1	0.34
	RMSE	0.27	0.24	0.45
	ACC	78%	83%	—

MAE:mean absolute error; RMSE: root mean squared error; ACC: accuracy

Table 3. Evaluation of classification and regression algorithms

attribute, while the second case augments sensor data with additional manually introduced data.

To compare how different learning methods behave on our dataset, we chose two classification algorithms: C4.5 algorithm for learning decision trees and Bayesian networks. For the regression method we applied a standard linear regression. We used implementation of these algorithms available on WEKA toolkit [2].

In Table 3 it can be observed how the algorithms performed on the simple and augmented dataset. The mean absolute error (MAE) and the root mean squared error (RMSE) are quantifying how close the predictions to the target value are. While MAE is an average of the absolute error, RMSE indicate square of the absolute error, emphasizing on how large is the difference between the predicted and actual values. For the classification algorithms, we also reported the classification accuracy.

As it can be seen in Table 3, all algorithms had better results with the augmented dataset, suggesting that combining sensor data with additional relevant data can help in improving crowdedness prediction. We can see that Bayesian networks have the best performance, closely followed by Decision trees (J48), while Linear regression has more than double error, compared to Bayesian networks on both datasets.

3.1. Decision tree

Figure 3 displays the decision trees for the simple and augmented datasets. For this representation, we put a constraint that at least 5% of the total number of instances has to be in a leaf (the constraint was set after several preliminary tests). Both decision trees have the sensor light measurements attribute in the root node, which can easily classify a large number of the instances with 0 persons, if the light has low value ($<= 25.42$). Another observation is that for the simple dataset, three of the four attributes are used for classification: light, temperature and humidity. Moreover, it can be observed how additional data influence the structure of the decision tree. In this case, only one attribute from the additional data is used: the number of computers working.

Considering that increasing the minimum number of instances in a leaf might generate data overfitting, but also the fact that other attributes

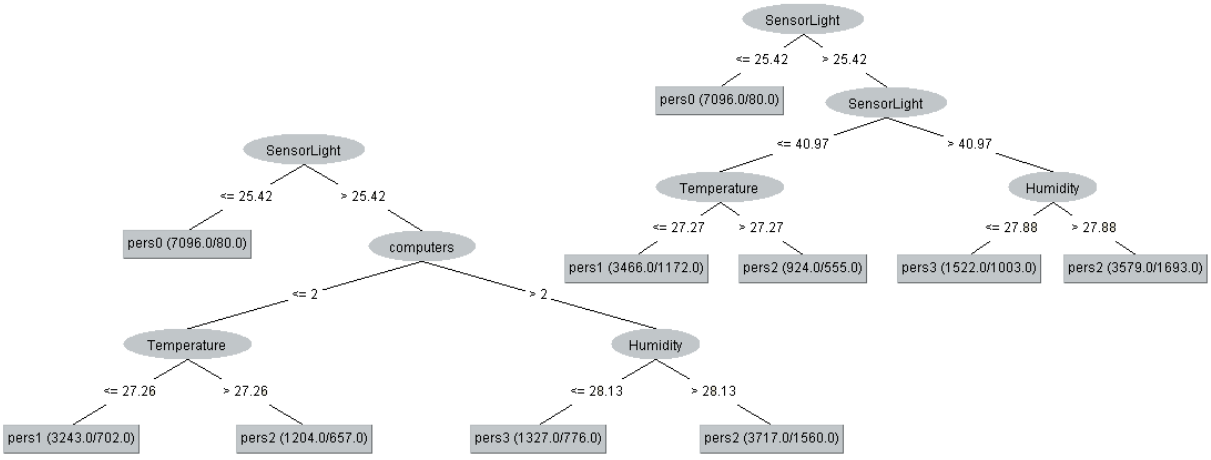

Figure 3. Decision tree for augmented (left) and simple (right) datasets.

might be useful for a better classification, we decided to continue our experiments with Bayesian network learning method, as this method describes probability distribution over a set of variables.

3.2. Bayesian network

For the Bayesian network algorithm we used the following setting in WEKA toolkit: simple estimator and K2 search algorithm.

Table 4 and Table 5 show the confusion matrix of the two cases of simple and augmented datasets. We can observe that in the case of the augmented dataset, there is a better distinction between the instances with 0 persons and the rest of instances. For example, there are no instances with 2 or 3 persons misclassified in the category with 0 persons.

	0	1	2	3
0	7096	120	100	11
1	95	2988	891	218
2	37	833	2566	253
3	1	246	492	640

Table 4. Confusion matrix for the simple dataset.

	0	1	2	3
0	7073	208	45	1
1	38	3085	933	173
2	0	557	2731	401
3	0	16	378	985

Table 5. Confusion matrix for the augmented dataset.

If we consider the class attribute as having only two values — 0 persons and more than 0 persons, then we can check if our system can be used for simple human detection. We performed a cost/benefit analysis and we were able to improve the accuracy of prediction on augmented dataset to up to 98%, compared to 83% correctly classified instances in the case of 4 categories (class attribute values). We chose to perform this analysis with Bayesian networks since it had the best overall performance. The results are shown in Table 6, where the cost and confusion matrix are represented. We set a 5 times bigger cost for misclassifying instances with more than 0 persons, then for misclassifying instances with 0 persons. As a result, there were only 26 instances with more than 0 persons misclassified, with a recall[1] of 0.997. This type of prediction can be useful in the case when it is more important to correctly classify the instances with more than 0 persons (e.g. alarm systems).

Cost matrix		**0**	**>0**	**Confusion matrix**	**0**	**>0**
	0	0.0	1.0		7042	285
	>0	5.0	0.0		26	9234
Acc.	98.12%					

Table 6. Cost/benefit analysis.

3.3. Linear regression

We have applied a simple linear regression algorithm. For the purpose of running linear regression, we need to have numeric values for all the attributes. Thus, the window attribute in the augmented dataset was mapped to three binary attributes (with values 0 or 1) corresponding to each nominal values (WinClosed, WinHalfOpen WinOpen). Also, the weekdays were mapped to a binary attribute (WorkingDay) with 0 for weekend and 1 for working days; the class attribute has integer values from 0 to 3.

Figure 4 depicts the resulting linear model used for prediction. It can be noticed that the same

People = 0.3913 * WorkingDay +
−0.0065 * Hours + 0.0537 * Temperature +
−0.009 * Humidity + 0.0105 * SensorLight +
−0.0059 * Pressure + 0.351 * Computers +
−0.6541 * WinClosed + -0.4602 * WinHalfOpen +
−0.5627 * WinOpen + 6.6509

People = 0.0482 * Temperature +
−0.0071 * Humidity + 0.0298 * SensorLight +
−0.0167 * Pressure + 17.0532

Figure 4. Linear regression model for augmented (top) and simple (bottom) datasets.

[1] Recall is defined as the ratio between the number of true positive instances and the sum of true positive and false negative instances.

attributes from the simple dataset are used as in the decision trees shown in Figure 3. We can see that temperature and sensor light have positive influence on the number of people. On the augmented data, the largest coefficients are at the number of computers and working day attributes, showing their positive influence on the predicted number of people. This indeed reflects the situation in the lab, as there are personal computers and each person usually turns off the computer before leaving the lab.

4. Related Work

Though the number of vertical systems implementations is not really high, similar work can be found in [3]. The authors present networked sensor infrastructure composed of commonly used devices in an office (PCs, PDAs, telephones etc.) to which a Bayesian ML method is applied in order to facilitate human interaction. The training set is composed by manually labeling each activity detected on the monitored devices. This approach is different from our work in terms of the utilized sensors and the accent is put on the Bayesian learning method, not on a vertical system.

The work presented in [4] is in the context of using semantic technologies in sensor networks. Using RDF and RDQL query languages with slight modifications, sensor data is modeled for querying in different situation. However, the dataset is obtained by simulating a sensor network that emphasizes the power of the system and query language, thus the system may perform differently in a real environment. A different approach is described in [5], where the sensor network is modeled using Dynamic Markov Random Field to analyze real-world environment, in which sensor data may be corrupted, influenced by noise or lost. Then the inference on data is done, using an implementation of two algorithms — Markov Chain Monte Carlo and Value Iteration — to predict and analyze forest fires.

5. Conclusions and Future Work

In this paper we presented vertical system integration for predicting the number of persons in our lab. We labeled sensor data with additional data and created an augmented dataset to which we applied ML algorithms. After analyzing the prediction results from the simple and the augmented datasets, we conclude that the number of persons can be predicted based on sensor data. Furthermore, the prediction can be improved when adding additional information for all three ML algorithms. In addition, we have also shown the improvements in accuracy of prediction when we limited the values of the class attribute to 0 or no persons.

Choosing the right ML method to apply on sensor data depends on the application and on the expected outcomes. On our data, decision trees and Bayesian networks give better results than linear regression, but, to make general conclusions, more experiments on larger datasets are needed. We found the model generated by decision trees to be the easiest to interpret and well performing.

The results we have obtained are encouraging for further extension of the system, by creating a network of sensors so that more information can be obtained. We are also considering complementing the current system with semantic technologies for enriching the data for more diversified and highly accurate predictions.

6. Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the European Community under PASCAL2 Network of Excellence (ICT-NoE-2008). This publication only reflects the authors' views.

References

[1] M. CERIOTTI, ET AL., Monitoring Heritage Buildings with Wireless Sensor Networks: The Torre Aquila Deployment. *Proceedings of the International Conference on Information Processing in Sensor Networks*, (2009) April 13–16, pp. 277–288, Washington, DC.

[2] M. HALL, E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, I. H. WITTEN, The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, (2009) Vol. 11, No. 1.

[3] M. MÜHLENBROCK, O. BRDICZKA, D. SNOWDON, J. MEUNIER Learning to Detect User Activity and Availability from a Variety of Sensor Data. In *Proceedings of the Second IEEE International Conference on Pervasive Computing and Communications*, (2004) March 14–17, Washington, DC.

[4] B. SZEKELY, E. TORRES, A semantic data collection model for sensor network applications. December 2004, `http://www.klinewoods.com/papers/semanticdcn.pdf` [02/15/2009]

[5] J.-M. KIM, Applying Dynamic Markov Random Fields for Sensor Data Analysis. Report, `http://web.mit.edu/murj/www/v13/v13-Reports/v13-r1.pdf` [02/15/2009]

Received: June, 2010
Accepted: November, 2010

Contact addresses:
Alexandra Moraru
J. Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: `alexandra.moraru@ijs.si`

Marko Pesko
MOBITEL Telecommunication Services Inc.
Vilharjeva 23, 1537 Ljubljana
Slovenia
e-mail: `marko.pesko@ijs.si`

Maria Porcius
J. Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: `maria.porcius@ijs.si`

Carolina Fortuna
J. Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: `carolina.fortuna@ijs.si`

Dunja Mladenic
J. Stefan Institute
Jamova 39, 1000 Ljubljana
Slovenia
e-mail: `dunja.mladenic@ijs.si`

ALEXANDRA MORARU is a student at the J. Stefan International Postgraduate School in the Information and Communication Technologies second-level program. She got her BSc degree in computer science from the Technical University of Cluj-Napoca in 2009. She started her collaboration with J. Stefan Institute in 2008, with a 2 months internship program, and since 2009 she is a student there. Her general research interests are in the area of Semantic Web and semantic technologies, more specifically, the applicability of semantics in sensor networks.

MARKO PESKO received B.Sc. degree in electrical engineering from the Faculty of Electrical Engineering, University of Ljubljana in 2009. He is employed at Mobitel, Telecommunication Services, Inc., the major Slovenian mobile provider, and is a second year Ph.D. student of information and communication technologies at the Jozef Stefan International Postgraduate School. His research work is carried out in the frame of young researcher scheme in collaboration with the Department of Communication Systems at the Jozef Stefan Institute. His main research interests are in the area of wireless sensor networks and their integration with next generation mobile networks.

MARIA PORCIUS is a student at the Jozef Stefan International Postgraduate School in the Information and Communication Technologies second-level program. She got her BSc degree in computer science from the Technical University of Cluj-Napoca in 2009. She started her collaboration with Jozef Stefan Institute in 2009, within Communication Systems department. She is also a member of SensorLab team, a group of mostly PhD students, who are developing their research work in the area of wireless sensor networks. Her research interests comprise challenges and issues on the network and application layers for sensor systems.

CAROLINA FORTUNA is a senior research assistant and a PhD student working at the Department of Communication Systems, Jozef Stefan Institute. She got her BSc degree in electrical engineering from the Technical University of Cluj-Napoca, Romania. Her research is interdisciplinary focusing on semantic technologies with applications in modelling of communication and sensor systems and on combining semantic technologies, statistical learning and networks for analyzing large datasets. She has published papers in refereed conferences and journals, served in the program committee of conferences such as ICC, Globecom and WCNC. She has also been to industry internships at Bloomberg LP and Siemens PSE.

DUNJA MLADENIC is an expert on study and development of machine learning, data/text mining, semantic technology techniques and their application on real-world problems. She has been associated with the J. Stefan Institute since 1987, first as a student and since 1992 employed as a researcher. She is leading Artificial Intelligence Laboratory of the Jožef Stefan Institute since 2011. She got her MSc and PhD degrees in Computer Science from the University of Ljubljana in 1995 and 1998 respectively. She was a visiting researcher at School of Computer Science, Carnegie Mellon University, USA in 1996–1997 and in 2000–2001. She has published papers in refereed conferences and journals, served in the program committee of international conferences and organized international events in the area of text mining, link analysis and data mining. She is co-editor of several books including "Data Mining and Decision Support: Integration and Collaboration", Kluwer Academic Publishers 2003, "Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies" Springer 2008, "Web Mining: from Web to Semantic Web", Springer 2004, "Semantics, Web and Mining" Springer 2006, "From Web to social Web: discovering and deploying user and content profiles", Springer 2007, "Knowledge Discovery Enhanced with Semantic and Social Information" , Springer 2009.