Collaborative Data Mining for Intelligent Home Appliances

Oliviu Matei, Giovanni Orio, Javad Jassbi, José Barata, Claudio Cenedese

Collaborative Data Mining for Intelligent Home Appliances

Oliviu Matei[1], Giovanni Di Orio[2], Javad Jassbi[2], José Barata[2], Claudio Cenedes[3]

[1] Technical University of Cluj-Napoca, Dept. of Elect. Eng, North University Center of Baia Mare, Baia Mare, Romania, oliviu.matei@holisun.com
[2] UNINOVA-CTS, Dept. of Elect. Eng, FCT-UNL, Caparica, Portugal
{gido, j.jassbi, jab}@uninova.pt
[3] Global Technology Center - GTC, Electrolux Spa, Pordenone, Italy
claudio.cenedes@electrolux.it

Abstract. The augmentation of physical devices and resources with electronics, software, sensing elements and network connectivity is a "hot topic" as confirmed also by the several research projects and activities on internet-of-things (IoT) and cyber-physical systems (CPS) research streams. It is obvious that intelligent products are taking more responsibility in future collaborative networks. Recent products are becoming more and more intelligent and connected by using the existing network infrastructure, meaning that products are becoming active agents in networks and valuable data sources that are capable to provide data continuously during their operation. This is leading to a massive amount of data that can be used by product manufacturers to be and remain competitive in market sharing. In this scenario, the application of collaborative data mining techniques, supported by machine learning algorithms, is aimed to enable the analysis of the data provided from multiple and above all distributed data sources in order to discover and extract useful knowledge about the behavior of the users along with the usage patterns of their devices and appliances.

Keywords: Collaborative Data mining, Intelligent Home Appliance, Collaborative Network

1 Introduction

Intelligent products are pushing more and more the Collaborative Networks (CN) discipline to a new level where the product itself is acting as an active and autonomous agent within an environment where people, processes, data, and things are connected together for designing/creating new products and services, while product manufacturers to more securely manage the risk and to extend the global reach by allowing higher collaboration from anywhere and at any time. In this scenario, the data continuously produced by intelligent products – if properly gathered and analyzed – can provide fundamental feedback information to product manufacturers about the products performance as well as products usage. This information can be used by the manufacturers for accelerating the product innovation process and increasing their competitiveness in

market sharing. For these reasons, products have become an indispensable part of New Generation of Collaborative networks ([1]).

According to Sanou [2], the access to the Internet has grown from an estimated 10 million people in 1993, to almost 40 million in 1995, to 670 million in 2002, and to 2.7 billion in 2013. As the connectivity and Internet of Things penetrate the daily life, a reasonable concern of the producers is to monitor and understand the behavior of the users along with the usage patterns of their devices and appliances. This would bring up invaluable knowledge as a basis for further technical and marketing developments, both in industrial and business environment, as well as in home context.

Therefore, much research has been invested in determine these usage patterns, such as the ones reported in [4] and [5]. Both use data mining to fulfil this objective. The aim is to determine an expected behavior of the device or user based on some known sensors outputs.

Human Agent collectives (HACs), are interesting investment for companies. Home appliance market leaders are keen to build smart houses (including smart kitchens) with the collaboration of both humans and intelligent products. This means HACs are the heart of smart collaborative networks while Intelligent Products are playing crucial role and the advantage is the stream of available data through sensors in the products. To manage this data/ Information in a productive way, different types of algorithm need to be employed. Collaborative Data Mining is a successful example of emerging algorithms to deal with "big data" collected from Collaborative Networks.

The term of collaborative data mining has been used several times, by Appleman in [6], Maimon in [7] and Moyle in [8]. But the term is not defined in a comprehensive manner that allows one to have a view of what are the advantages of such an approach compared to classical data mining, what are the benefits of such an approach, what are the underlying principles of such an approach and how this approach may be different compared to other approaches. Later on, in 2008, Zhan et al. [9] used the term collaborative as referring to using more data sources for mining.

As the connectivity is present in our daily lives [10], the quantity of available data is ever increasing and companies have real interest in understanding the behaviour of their customers [11], a trend is to collaborate all these data source for more accurate analysis [12]. The research questions are:

- *What is the influence of using the data from two sources to predict the behavior of one source?*
- *Is there any relationship between the accuracy of such predictions and the correlation between the two sources?*

2 Collaborative Data Mining

As stated in [13], collaborative data mining is a setting where the data mining effort and/or process is distributed among multiple collaborative agents that are autonomously executing actions in the environment. Collaborative data mining is based on the main assumption that collaborative data mining processes can better tackle the considered data mining process with respect to a data mining process built on the top of individual

and non-collaborative agents. Data mining is the process of extracting information or better knowledge from large quantity of data. Collaboration is the act of a variety of entities to share information, resources and responsibilities to jointly plan, implement, and evaluate a program of activities to achieve a common goal [14]. As stated in [14], collaboration is a difficult process and typically its success depends on several requirements, namely:

i) the main purpose for collaboration;
ii) the initial baseline for collaboration;
iii) the collaborative process (a set of generic steps);
iv) the creation of a space for collaboration;
v) and finally the definition of resources, rewards, commitment and responsibilities.

Despite these difficulties the main motivating point is that by using collaborative processes it is possible to reach results that cannot be reached by parties/entities/agents working alone. With this in mind, the proposed research is based on the assumption that there are similarities between usage patterns of the same devices or more in general agents [2]. In such a way, we propose a data mining model which takes into account also the behavior of other similar devices/agents in the region when doing patterns recognition. This extends the individual data mining with the input set of other devices/agents, while meaning that the size of the input set doubles with the data from the sensors of corresponding devices as shown in Table 1

In the Table 1, the first 5 columns are the data from the devices/agents to be analyzed, namely the timestamp, from which the date of the week ("DW"), the date and the hour of the day ("Hr") are derived. The column "Door_close" shows the event of opening or closing the door, not whether or not the door is opened. That is why two successive records always have a true and a false value for "Door_close". The same data (shown in the main column "Similar device") is available for another device. While the classical data mining tries to predict the outcome as a Boolean value (column "Door_close") only based on the data from the same device/agent (main column "Analyzed Device"), the collaborative data mining take into account all available data, in this case both columns "Analyzed Device" and "Similar device".

Thus, in the context of this paper collaborative data mining is intended as the application of data mining techniques on multiple data sources (autonomous agents) to better understand the behavior of one of them as well as the correlation – if it exists – between them. The rapid progress on computer networks and the pervasive computing – supported by new computing paradigms and technologies such as IoT, CPS, web services (WS), cloud computing, Service Oriented Computing (SOC), etc – is offering the technical infrastructure and foundation for new forms of collaboration in several context of application [15], [16], and where the proposed model can be easily deployed. In particular, for this research we have decided to use the technology and the infrastructure provided under the scope of EU-FP7 ProSEco1 project, where intelligent products (intelligent autonomous agents) are providing data during their operation. This data is then used for applying the proposed collaborative data mining process i.e. as input to the

[1] https://www.proseco-project.eu

data mining process. More specifically, we aim to combine the results generated by data mining processes – applied on isolated agents (i.e. devices) – in a collaborative manner, or better, the usage of the obtained results in order to refine and improve the outcome of the data mining process on a selected agent (see Fig. 1).

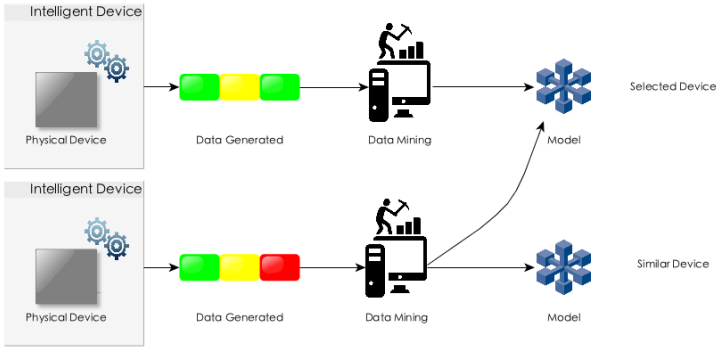

Fig. 1. Proposed collaborative data mining model flowchart

Table 1. Some samples of the data set in the case of collaborative data mining

Analyzed Device					**Similar device**				**Out-come**
Time-stamp	**DW**	**Date**	**Hr**	**Door_ Close**	**Time-stamp**	**DW**	**Date**	**Hr**	**Door_ Close**
15.11.2013 00:28	**6**	**15.11.2013**	**0**	T	15.11.2013 00:49	6	15.11.2013	0	**T**
15.11.2013 00:28	**6**	**15.11.2013**	**0**	F	15.11.2013 00:49	6	15.11.2013	0	**F**
15.11.2013 01:11	**6**	**15.11.2013**	**1**	T	15.11.2013 02:15	6	15.11.2013	2	**T**
15.11.2013 01:11	**6**	**15.11.2013**	**1**	F	15.11.2013 02:15	6	15.11.2013	2	**F**
15.11.2013 01:40	**6**	**15.11.2013**	**1**	T	15.11.2013 02:45	6	15.11.2013	2	**T**
15.11.2013 01:41	**6**	**15.11.2013**	**1**	F	15.11.2013 02:45	6	15.11.2013	2	**F**
15.11.2013 01:44	**6**	**15.11.2013**	**1**	T	15.11.2013 03:18	6	15.11.2013	3	**T**

3 Methodology

We aim to compare the collaborative data mining with standalone data mining, hence we use the same research methodology as reported in [4]. The training data records are structured like in Table 1.

For collaboration, three devices have been selected randomly, along with the one used in [3] and [4]. The appliances are real ones, deployed by Electrolux to real customer in the United States, for tests. The monitored period is 13th of November 2013 - 25th of December 2013, with 1170 data records.

The objective is to determine the number of door openings per day. This process aims to determine how many times the device will be opened based on the number of openings in the previous days. This would be a good statistical indicator about the usage of the device and could help to manage internal functions (such as "Defrosting" procedure of refrigerators) in an efficient way. As the records are event-based, the data is aggregated over the day. The number of openings is a count of all records for which the value of "Door_close" is T (true).

The experiments have been done for the following benchmarks:

1. oven 1, 13, respectively oven 14 alone;
2. oven 13 using the previous data for oven 14, respectively oven 14 using the previous data for oven 13. The two ovens have been chosen because there is a high correlation between them. The mathematical foundations of the correlation are explained in section 3.1.
3. oven 13 using the previous data for oven 11, respectively oven 1 using the previous data for oven 13. Between oven 1 and oven 13 there is a very low correlation (0.01).

The first set of benchmarks would provide a standard for comparing the results of the benchmark sets 2 and 3.

The methodology refers to the correlations between some ovens. The definition and the computations of the correlations are shown in the next section.

3.1 Correlations between the Benchmark Appliances

Correlation between some benchmark appliances means correlation between some time series consisting of the usage patterns of the devices over the time [17]. The main challenge here is that the two series do not have the same time samples, although the tick is the same, e.g. one can be sampled at 15.11.2013 11:30:31; 15.11.2013 11:30:31 and the other one 15.11.2013 11:32:45; 15.11.2013 11:42:45 etc. Another challenge is that the two series may be correlated lagged in time. In this case, cross-correlation is the best function to be computed, according to Aarts et al. [18]. Considering two series *x(i)* and *y(i)*, where *i=0, 1, 2,...N-1*, Bourke [19] defines the cross correlation *r* at delay *d* as:

$$r(x,y) = \frac{\sum_i [(\mathrm{x(i)} - \mathrm{mx})(\mathrm{y(i} - \mathrm{d}) - \mathrm{my})]}{\sqrt{\sum_i (\mathrm{x(i)} - \mathrm{mx})^2}\sqrt{\sum_i (\mathrm{y(i)} - \mathrm{my})^2}} \quad (1)$$

where *mx* and *my* are the means of the corresponding series.

Out of the 143 assets monitored by Electrolux, for which data is collected, only the ovens Kenmore Elite 97102 (counting up to 14 pieces) have the "Door_close" event, which is used in this research. Therefore we cannot talk about other correlations expect between the usage of these. Fig. 2 displays graphically the collaborative network of the

ovens over a period when most of the ovens were used, namely December 2013. The nodes represent the ovens, denoted as numbers for a better comprehension. Their size is proportional with the total correlation for that respective oven, computed as:

$$r(i) = \sum_{i} \mathrm{r(i,j)} \tag{2}$$

The oven number "2" misses because that oven starts being used only in April 2014.

The vertices are inverse proportional with the correlation of usage of the respective ovens. The closer the nodes, the more cross-correlated the ovens. The cross-correlation factor is in absolute values, as the negative values represent a negative relationship (yet existent). Ovens "9" and "14" have a very similar behavior, which is why they overlap. On opposite, the farther the nodes (or the longer the vertex), the lower correlation between the usages of the respective ovens. If there is no vertex between two nodes, it means that the correlation between the nodes is almost zero and was not displayed in the network.

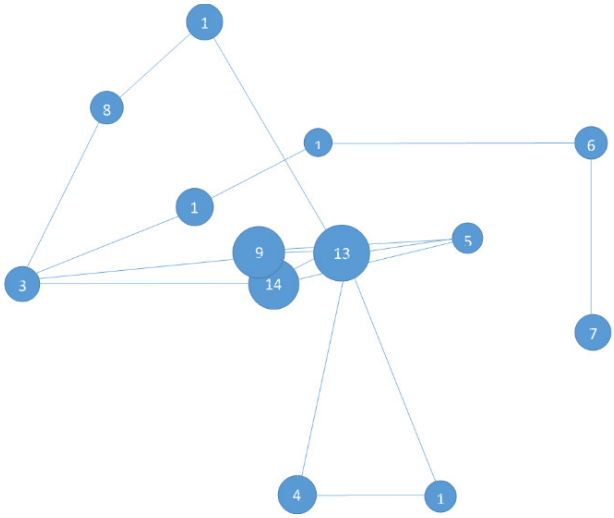

Fig. 2. Collaboration network based on the correlation between the usages patterns of the ovens

Fig. 3 shows the correlations between usage patterns of the ovens over the whole period (November 2013 - April 2014) as a bubble chart. The larger the diameter of the bubbles, the higher the correlation. The smaller the bubbles, the lower the correlation. Please not that the positive and negative correlations are represented having the same diameters. The chart is symmetric as the correlation is symmetric. A missing bubble, like for the pairs (2, 4), (2, 5), (2, 6) and so on, means that the correlation is between the usage of the respective ovens is zero. It is the case of the oven 2 because it was used for a short period of time.

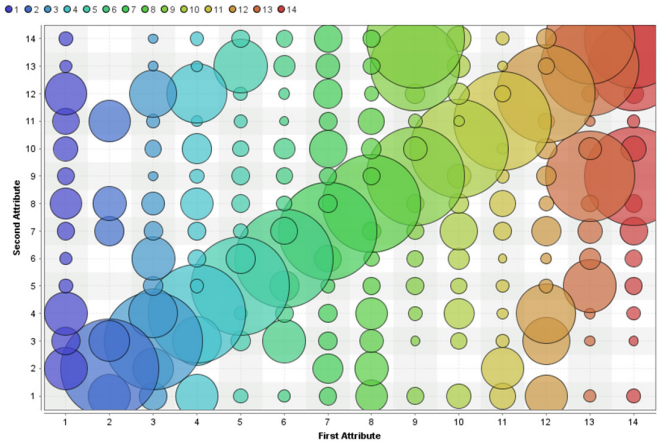

Fig. 3. The correlation between usage patterns over the whole period

The experimental results are summarized in the next section.

4 Experimental Results

Three sets of benchmarks have been experimented, as shown in section 3:

- For three ovens (1, 13 and 14) independently;
- For oven 13, respectively 14, using the data from both 13 and 14, as the usages of the two ovens are correlated;
- For oven 1, respectively oven 13, collaboratively, as the usages of the two are not correlated.

The data mining tool is Rapid Miner because it provides also the libraries with which it can be embedded in any Java software. Therefore it can be used also in real field applications, as this research is meant to be a first step for introducing intelligent features such as predictive maintenance in electronic appliances.

Several algorithms have been used: LPR (local polynomial regression), SVM (support vector machine), NN (neural network) and k-NN (k-nearest neighbour). The results are summarized in Table 2. For avoiding overtraining, 20% of the data was set aside for cross-validation.

The results are summarized in Table 2 for ovens 13 and 14, respectively in Table 3 for ovens 1 and 13. The first column displays the algorithm used, as the ones reported in [3]. The next two columns present the accuracy for each oven in the case of individual data mining. The last two columns present the accuracies for predicting the number of openings for each oven when applying collaborative data mining.

When the usage patterns of the ovens are correlated (see ovens 13 and 14, summarized in Table 2), the collaborative data mining got the best results (highest accuracies). In one case, the stand-alone and collaborative approaches had the same accuracy for

oven 13 (in the case of SVM). For oven 14, in two cases the results were not improved by collaborative data mining (algorithms SVM and NN). However, the average accuracies for collaborative data mining are better.

Table 2. The accuracies for stand-alone, respectively collaborative data mining of ovens 13 and 14 (highly correlated in terms of usage patterns)

Algorithm	Window size	Oven 13 alone	Oven 14 alone	Oven 13 collaborative	Oven 14 collaborative
LPR	3	81.5%	91.3%	**80.0%**	**92.0%**
SVM	5	**91.3%**	**95.7%**	**91.3%**	**95.7%**
NN	1	46.2%	**51.9%**	**48.1%**	**51.9%**
k-NN	3	80.4%	96.0%	**92.6%**	**96.2%**
Average		**74.85%**	**83.73%**	**78%**	**83.95%**
Improvement				**4.21%**	**0.27%**

The second researched case is of collaborative data mining when there is a very low correlation between two ovens, in this case, ovens 1 and 13, with the correlation factor of 0.01. The experimental conditions are similar with the previous benchmark. In this case, the collaborative accuracies are lower than stand-alone ones with 8.91% for the oven 1, respectively 1.4% for oven 13 (see Table 3).

Table 3. The accuracies for stand-alone, respectively collaborative data mining of ovens 1 and 13 (loosely correlated in terms of usage patterns)

Algorithm	Window size	Oven 1 alone	Oven 13 alone	Oven 1 collaborative	Oven 13 collaborative
LPR	3	**48.1%**	**81.5%**	32.0%	70.6%
SVM	5	81.5%	**91.3%**	**91.3%**	**91.3%**
NN	1	**18.5%**	46.2%	14.8%	**48.1%**
k-NN	3	**36.0%**	80.4%	29.6%	**85.2%**
Average		**46.03%**	**74.85%**	**41.93%**	**78%**
Improvement				**-8.91%**	**-1.4%**

The improvements in accuracy of collaborative data mining has proven reasonable, but including context information, such as outside temperature and number of users could increase the accuracy even more. However, this kind of data is not available for now.

5 Conclusions and Discussions

We have shown that collaborative data mining means the use of data from two different data sources for determining the output of one source (in this case home appliances, but the specific embodiment of the source has no significance). We have proven that if the two sources are correlated, the accuracy of the data mining or machine learning process increases, whereas if the two sources are not correlated, the accuracy decreases. Eventually this could lead to predicting the output of a source using only the data from another, correlated source when the data from the former one is not available.

The main challenge of this work is dealing with information collecting from distributed system which would be recognized as Collaborative network. This information provided by real business cases from Electrolux using intelligent products. We have proven that there are varying similarities between usage patterns of the 143 electric appliances (refrigerators and ovens) in the field, studied in our research. Of course, those similarities influence the impact on the data mining results. This is why the average values displayed in Table 2 are quasi-equal, but the best values are significantly better in the case of collaborative data mining. Of course, those values are obtained when the data mining process takes into account devices with similar usage patterns.

This concept of "collaborative data mining", can be applied in any case when there are more devices with similar behavior which could be ideal tool dealing with huge amount of information in intelligent Products Ecosystem. Therefore, before using it, a correlation analysis must be performed. Such approach fits very well in the context of Industrial Internet (see also [4]), where the number of machines are relatively low, but with similar behavior. This could help to analyze distributed information from intelligent collaborative products.

6 Further Developments and Extensions

It is beyond the scope of this paper to research what is the (mathematical) relationship between the accuracies in the collaborative data mining and the series correlations described in Section 5.

Further research needs to be carried out on collaborative data mining with at least three devices, with various correlations. Another topic is how a series alone influences another series, relatively to the correlation factor. And yet, how can be handled the common case when the two series are not aligned temporarily.

As principle, collaborative data mining can be applied in any field, with any data arguments. However, the heuristics are obviously specific for each case, as well as the tuning of the parameters of the machine learning algorithm. Also extending the proposed methodology for analyzing the behavior of HACs could be interesting direction for future researches. Finally, from a more technical point of view further research is needed on identify a common representation for the data coming from the collaborative agents. As a matter of fact, as the number of collaborative agents are increasing more and more it is necessary to find a common way to represent the data in order to allow the data mining process to analyze get the data always in the same way, i.e. with the

same syntax and semantic. It is a necessary issue that need to be considered and handled whenever new type of agents are considered and new context of application are studied.

Acknowledgments. This work is partly supported by the ProSEco project of EU's 7th FP, under the grant agreement no. NMP-2013 609143.

7 References

1. W. Picard, "Resilient and Robust Human-Agent Collectives: A Network Perspective," in Risks and Resilience of Collaborative Networks, L. M. Camarinha-Matos, F. Bénaben, and W. Picard, Eds. Springer International Publishing, 2015, pp. 79–87.
2. Brahima Sanou, "ITC Facts and Figures 2013", Telecommunication Development Bureau, International Telecommunications Union (ITU), Geneva, Feb 2013. Retrieved 23 May 2015.
3. Oliviu Matei, Kevin Nagorny and Karsten Stoebener, "Applying data mining in the context of Industrial Internet" International Journal of Advanced Computer Science and Applications(IJACSA), 7(1), 2016. http://dx.doi.org/10.14569/IJACSA.2016.070184
4. Giovanni Di Orio, Oliviu Matei, Sebastian Scholze, Dragan Stokic, J. Barata, Claudio Cenedese, A Platform to Support the Product Servitization, IJACSA, vol. 7, no. 2, 2016
5. Oliviu Matei, PRELIMINARY RESULTS OF THE ANALYSIS OF FIELD DATA FROM OVENS, Carpathian Journal of Electrical Engineering, vol. 8, no. 1, 2014
6. Appleman, Kenneth H., et al. "Collaborative internet data mining systems." U.S. Patent No. 5,918,010. 29 Jun. 1999.
7. Maimon, Oded, and Lior Rokach, eds. Data mining and knowledge discovery handbook. Vol. 2. New York: Springer, 2005.
8. Moyle, Steve. "Collaborative data mining." Data Mining and Knowledge Discovery Handbook. Springer US, 2009. 1029-1039.
9. Zhan, Justin. "Privacy-preserving collaborative data mining." Computational Intelligence Magazine, IEEE 3.2 (2008): 31-41.
10. Haythornthwaite, Caroline. "Social networks and Internet connectivity effects." Information, Community & Society 8.2 (2005): 125-147.
11. Heierman III, Edwin O., and Diane J. Cook. "Improving home automation by discovering regularly occurring device usage patterns." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.
12. Hajibandeh, N., et al. "Resemblance measurement of electricity market behavior based on a data distribution model." *Int J. Electrical Power & Energy Systems* 78 (2016): 547-554.
13. S. Moyle, "Collaborative Data Mining," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds. Springer US, 2009, pp. 1029–1039.
14. L. M. Camarinha-Matos and H. Afsarmanesh, "Collaboration forms," in *Collaborative Networks: Reference Modeling*, Springer US, 2008, pp. 51–66.
15. L. M. Camarinha-Matos, H. Afsarmanesh, "Motivation for a theoretical foundation for collaborative networks," in *Collaborative Networks: Reference Modeling*, Springer, 2008, pp. 5–14.
16. G. D. Orio, D. Barata, A. Rocha, and J. Barata, "A Cloud-Based Infrastructure to Support Manufacturing Resources Composition," in *Technological Innovation for Cloud-Based Engineering Systems*, Springer, 2015, pp. 82–89.
17. Wei, William Wu-Shyong. Time series analysis. Reading: Addison-Wesley publ, 1994.
18. Aarts, Ronald M., Roy Irwan, and Augustus JEM Janssen. "Efficient tracking of the cross-correlation coefficient." Speech and Audio Processing, IEEE Trans. on 10.6 (2002): 391-402.
19. Bourke, Paul. "Cross correlation." Cross Correlation", Auto Correlation—2D Pattern Identification (1996).