PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

Ph.D. Student: Vincenzo Schiano Di Cola

XXIX Cycle

Training and Research Activities Report – First Year

Tutors: Nicola Mazzocca, Francesco Piccialli

1. Information

This report is written by Vincenzo Schiano Di Cola. I received my master’s degree in Mathematics from this same University. As the reader can infer from the header, I belong to the XXIX Cycle of the ITEE at the Università di Napoli Federico II, and the tutors who are supervising me are Prof. Nicola Mazzocca and Prof. Francesco Piccialli.

I received a regional fellowship, the P.O.R. entitled: *Data Scientist for Predictive Analytics*; according to the "*Dottorati di ricerca con Caratterizzazione Industriale*" - DGR n. 156 del 21/03/2017 - DD n. 155 del 17/05/2018 - within the *POR Campania FSE 2014/2020 - Obiettivo Specifico 14 - Azione 10.4.5*.

2. Study and Training activities

During this first year, I gathered a number of credits summing up, roughly, to 35 ECTS. They are distributed among seminars and courses, both taken either aboard and within Naples.

These activities are here listed, with their title, type and number of credits:

- L’Accademia delle Startup - Le Startup dell'Accademia; Module: Improvement of research skills; 0,8
- A leap into Functional Data Analysis: from theory to applications; Seminar; 2
- Advanced techniques for software robustness and security testing; Ad hoc module; 3
- Data Science and Optimization; Ad hoc module; 1,2
- Computational and Machine Learning Methods for Complex Ecosystems; Seminar; 0,2
- Chaos in Magnetization Dynamics; Seminar; 0,3
- Advanced techniques for image denoising; Ad hoc module (Uniparthenope); 3,2
- Intelligenza Artificiale; MS Module; 6
- Big Data Analytics & Business Intelligence; MS Module; 6
- Non-Negative Matrix Factorization; Seminar; 0,4
- PCA, LDA and LVQ Classifiers; Seminar; 0,6
- eBISS 2019; Summer School (Germany); 4,8
- Lipari School 2019: Data Science; Summer School ; 6
- Advanced technology at the service of visitors to cultural heritage sites; Seminar; 0,4

3. Research activity

The research project involves three main participants that enrich my research with their specific objectives and expertise, within the more general framework of Data Science. These three members of the POR fellowship are the DIETI at University of Naples “Federico II”, the IT company DATABOOZ ITALIA S.r.l., and the Chung-Ang University in South Korea.

Up until now, the overall research output has focused on the use of Machine Learning techniques for IoT in the Cultural Heritage domain and unsupervised classification of mixed data: time (numerical) and path (strings). These research topics are matched and planned, also with the IT company DATABOOZ ITALIA S.r.l. in the person of Ing. Paolo Benedusi.

The research questions I tried to answer were:

1) Which unsupervised method is more suitable for studying visitors' paths in a museum?
2) How to select the correct number of clusters behaviors?
3) How to correctly assess the quality of a cluster?
4) What are the correct distances to use, when dealing with non-numerical features like paths?
5) What insights can such clustering of visitor's behavior give to museums' decision-makers?

Coherently with the nature of Data Science, as the intersection of math, IT and domain knowledge, most of the collaborations has been pursued with the Department of Mathematics and Applications (DMA) "Renato Caccioppoli", and with CINI, the Consorzio Interuniversitario Nazionale per l'Informatica; although future research, hopefully, is to be done also with other foreign universities.

The research of this first year is mainly shown in the listed papers in Section 4. Here is a brief description.

Classification techniques are applied to two Cultural Heritage domains. First, an IoT dataset of visitors inside the National Archaeological Museum of Naples (MANN), collected through a smartboard with Bluetooth capabilities. Next, the log usage of how users interact, on a tablet, with artworks linked to IoT Beacon sensors located in a castle named Castel Nuovo (also known as Maschio Angioino). The experience in a cultural space advises us to distinguish visitors in a series of categories since each visitor carries an invisible label listing which desires, expectations and needs relate to the experience inside a museum, referable to different types of visitors.
The main research goal is to extract knowledge from the available IoT data; in particular, it has been explored a kind of dataset composed of mixed data (numerical and non-numerical data). The results are mainly on searching for a suitable number of clusters in the unsupervised learning framework, and explaining he obtained clusters within the museum domain. Given a collection of paths, expressed as a sequence of letters in a string, e.g. XIRX. Where each letter indicates the location where a visitor has been located by a Bluetooth receiver, and the total time length of the stay, e.g. 1:02:32. The first main problem is to understand how two paths (i.e. strings) are similar. Paper 1. "A machine learning approach for IoT cultural data", addresses this problem. To deal with the non-numerical Path feature of our dataset, that can be treated as a string, we have considered four strings similarities and distance functions. In order to look at the problem by different perspectives, we used either edit-based distances (generalized Levenshtein, Longest Common Substring) and q-gram based distances (Jaccard and cosine). In order to evaluate clustering results, it has been selected the silhouette index extending, in Paper 2., to five different kinds of indexes: the SSE (sum square error), McClain-Rao, C-index, Silhouette, and Dunn index.
To choose the number of clusters, k, a heuristic approach is used. Based on the suggested methodologies implemented in the NbClust package and customizing the final selection. With the aim of identifying a prominent elbow suggesting the most appropriate number of existing clusters inside the data, i.e. difference-like criteria, and similar optimization-like criteria have been applied. In particular, instead of focusing only on two consecutive data partitions (formed by k and k+1 clusters), we looked at the triple (k-1, k, k+1). To select the final k, an ensemble approach through a voting technique has been used. Each

Hierarchical Clustering (with a specified linkage-distance method), has been evaluated with 5 different indexes and consequently, each cluster-index expresses its best k value. Finally, to cluster the data and be less sensitive to outliers, we used the PAM algorithm.
Paper 1., analyses the resulting cluster, with comparisons between several similarities and distance functions that suggest us a useful hidden path in the data. In the following work of Paper 2., we considered the time feature in order to improve the classification approach and generalize the comparison by obtaining an ordered F1 Score heatmap representing the comparison among all the experimented clustering methodologies. And presented a Three Fields Plot representing a comparison between different clusters with respect to time quantiles.
Finally, Paper 3. on "Decision Making" uses the results of the previous papers, to give meaning at the clusters, and propose insights for museum decision-makers. Paper 3. present the MANN case study, that has been extensively studied in Paper 1. and 2., and results on the Castel Nuovo museum, whose data has more multimedia feature, such as the number of actions interrupting an audio track or number of actions on photos. For the Castel Nuovo dataset, we try to check whether or not the use of technology can lead to homogeneous behaviors among different visitors interacting with mobile devices. A paper under review discusses the process of obtaining clusters from the Caste Nuovo data.

Further research in this area of path analysis involves understanding Sequential Decisions via Inverse Reinforcement Learning.

The most recent data to analyze in my research is on medical data. This data is > 18Gb and is labeled. So the idea, in this case, is to apply supervised learning. The main techniques to study are decision trees (random forest), KNN (k-nearest neighbor algorithm), SOM (Self Organizing Maps), LDA, and Deep learning techniques.
Moreover considering the complex interconnections in this new dataset, research will involve knowledge graphs, ontologies, graph databases, and models for relational learning, knowledge graph embeddings and link predictions.

4. Products

This first year of research lead to three publications:

1. "A machine learning approach for IoT cultural data" on the *Journal of Ambient Intelligence and Humanized Computing*. First Online: 04 September 2019. DOI: 10.1007/s12652-019-01452-6
2. "Exploring Unsupervised Learning techniques for the Internet of Things" on *IEEE Transactions on Industrial Informatics*. Date of Publication: 12 September 2019. DOI: 10.1109/TII.2019.2941142
3. "Decision Making in IoT Environment through Unsupervised Learning" on *IEEE Intelligent Systems*. Date of Publication: 01 October 2019. DOI: 10.1109/MIS.2019.2944783

Others are in preparation. One paper is under review at Springer's Journal *Multimedia Tools and Applications*, which is about "Unsupervised learning on multimedia data". Further papers will, eventually, be on the most recent research developments.

As for patents, none have been developed. Moreover, still, some analysis has to be done, in order to understand the consequences of what to *patent* in a Data Science framework.

5. Conferences and Seminars

As for this point, I did not present my work at a conference but, during this first year, I had the opportunity to present my work at a poster session and at a mini-symposium, both in July.

The poster session was held during the eBISS 2019, from the 1st of July to the 4th of July in Berlin, Germany. Together with Dr. G.Casolla, we presented two posters: “Unsupervised Learning: Similarities and Distance Functions for IoT Data” and “Unsupervised Learning: A Time Perspective Analysis of Visitors' Behaviors”.

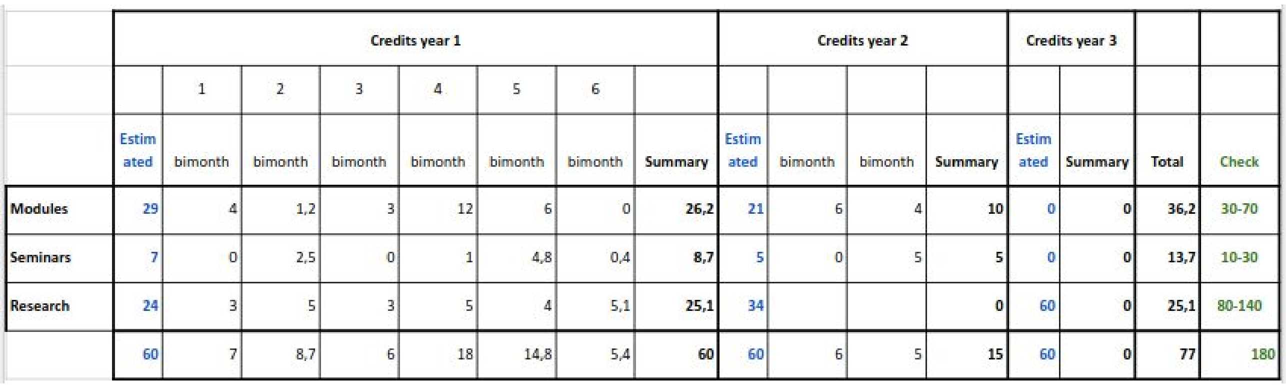
A week after, on day 11/07/2019 during the Young researchers mini-symposium, at the INDAM Intensive Period 2019, I held a presentation entitled: “*Partitionings and Similarity Metrics in Unsupervised Learning*”.

6. Activity abroad

For this first year, there has been no activity aboard. However, during the various seminars, and Summer School I had the chance to meet other researchers, especially Ph.D. students in the area of Data Science. These contacts might lead to some collaborations.

7. Tutorship

No mentoring has been done, since my POR fellowship is on a tight schedule, and each hour is reported and accounted for the Campania Region.



	Credits year 1								Credits year 2				Credits year 3			
		1	2	3	4	5	6									
	Estimated	bimonth	bimonth	bimonth	bimonth	bimonth	bimonth	Summary	Estimated	bimonth	bimonth	Summary	Estimated	Summary	Total	Check
Modules	29	4	1,2	3	12	6	0	26,2	21	6	4	10	0	0	36,2	30-70
Seminars	7	0	2,5	0	1	4,8	0,4	8,7	5	0	5	5	0	0	13,7	10-30
Research	24	3	5	3	5	4	5,1	25,1	34			0	60	0	25,1	80-140
	60	7	8,7	6	18	14,8	5,4	60	60	6	5	15	60	0	77	180