Social mining from Wi-Fi campus data

Roberto Puccetti [1]
supervised by Dino Pedreschi[1] - Mirco Nanni [2]

[1] Pisa University, Pisa - Italy
[2] CNR Pisa – Italy
roberto.puccetti@unipi.it

Abstract. This research deals with methods and algorithms for analyzing Wi-Fi data of individuals at the scale of a large campus area, studying personality traits, class attendance, mobility and social network structure of students. In particular, the main driving application will be the study of whether and how such elements influence academic performances.

1 Introduction

Data mining has been used in telecommunication industry for several applications including marketing, security and network reliability. The exploration data in educational field using Data Mining techniques concerns with extracting a pattern to discover hidden information from educational data. In the case of academic performance studies, a number of behavioral patterns have been linked such as time allocation, active social ties, sleep duration and sleep quality, or participation in sport activity. While most of the existing studies suffer from biases and limitations often associated with surveys and self-reports, our research is directed towards a Wi-Fi network in an urban campus area and therefore it analyses a large set of data that are not biases influenced. We want first to address the problem to analyze such data to discover their quality and then to develop a tool to enable the extraction of latent knowledge in dynamic and multidimensional networks. As an application area of the studies, we use the movements within a university campus and then the study of the influence of daily behavior on students' performance.

2 Related Works

We can define our research as multidisciplinary, in the sense that it approaches and uses many fields of information science: Wi-Fi Data Analysis, Social Networks (Construction [1], Topological properties [2] and Study of community [3]) and Privacy issues [4]. However the most of survey work done is about Prediction of Students' Academic Performance [5] [6], [7] [8], [9], [10]. In these works there are many aspect similar to our research but the most innovative aspects we have introduced are: no survey from questionnaires but directly from their behavior using a large dataset collected from students (about 50k) by their tab-

SEBD 2019, June 16-19, 2019, Castiglione della Pescaia, Italy.

let/smartphones, plurality in course of study, and Social network observations from campus Wi-Fi data (their visited places and physical proximity at the same time).

3 Research plan

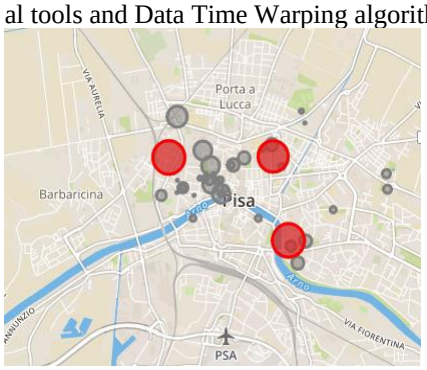
To gain the goal of the research, activity includes the following main tasks: to adopt or develop methods and algorithms to analyze wi-fi data on the scale of a large campus, to develop a software prototype, on top of such methods and algorithms, that can be used for solving a selected set of problems as, for example, to predict students' academic performance. Finally, we want to collect a "so-big dataset" from a campus area as a case study to test the prototype. From the viewpoint of the case study, we are planning: i) to extract and evaluate the importance of different sets of features for supervised learning models in particular for students' performance prediction; ii) to identify individual and network factors that best correlate with students' performances; iii) to predict students' performance; iv) to investigate significant differences among performance groups, in terms of the most important individual and network features. Data used to build the case study come from a consistent dataset: Wi-Fi access logs and exam results from the University of Pisa, described in better detail later.

4 Preliminary results

The Wi-fi data usually suffer from various issues, such as sparse data, noise, uncertainty and so on, which need to be dealt with before any analysis-based task. We start our work by studying and trying to remove (or to mitigate) them.

4.1 Semantic labelling of Access Points

Another preliminary work is to understand the role of each Access Point connection in the people activity. This means to understand two main important features of the access points: collocation in area and purpose of their use. So, we aggregated APs to calculate and display interactively the daily use (see Fig.1) and we identified 11 classes of use (Didactics, Central Administration, Study area, Recreational activities, Dormitory, etc.). This second task has two approaches: *Top-Down*, supervised (based on collocation, and *Data-Driven*, using visual tools and Data Time Warping algorithm to validate what prefixed in arbitrary way.



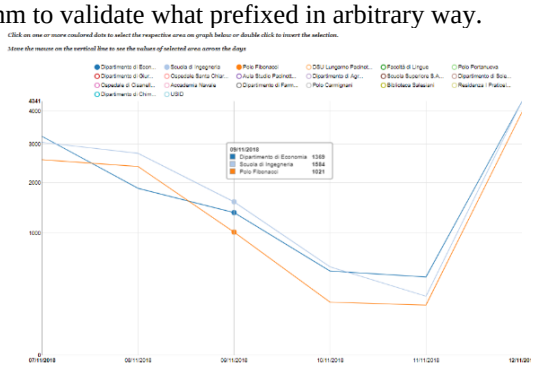


Fig. 1. The 3 main used areas (from http://sintetik.altervista.org/visual/)

4.2 Students clustering

We use two different clustering methods to study users' ties: **behavior similarity**, that wants to analyze how much students use APs in percentage and for which purpose, and **physical collocation**, that presupposes the friendship between two individuals who habitually attend the same places. The first, for each AP classification gathers a score from 1 to 2 to "weight" the importance of the location frequency, for instance assigning higher weights to spaces devoted to study and (secondarily) social activities. Then, for each category, each student is represented by an array describing the percentage of his time spent in it. This representation of a student allows the direct comparison of two individuals through a simple weighted sum (basically a L1-norm where each feature has an associated importance). Distance from student *i* to student *j* is

$$d(i,j) = \sum_{n=1}^{K} abs(V_{i_n} - V_{j_n}) * W_n$$

$k = number\ of\ item\ V_{i_n} = Value\ of\ item\ n\ in\ array\ i \quad W_n = weight\ of\ item\ n$

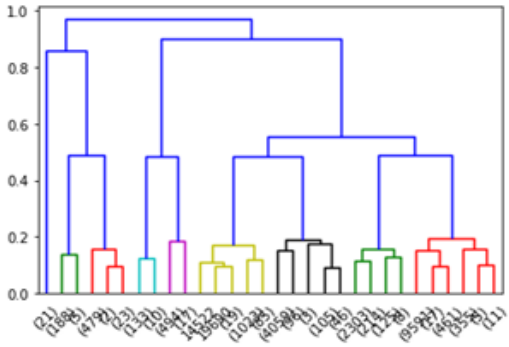


Fig. 2. Hierarchical clusters dendrogram of students' similarity

In such a way, we create a student's *weighted disctance matrix* to measure distance for clustering operations, aimed to group students that spend their time in a similar way (Fig.2).

Instead, the physical location approach considers the simultaneous use of the access point as similarity factor. Staring from the time logs, we study the overlapping time in function of the APs use to determine the reasonable cut factor. We decided to consider 20 minutes as reasonable overlapping time. With this parameter and adopting Jaccard distance index, we construct the relationship matrix for social network methods to determine the student's community. Using two different methods (DEMON and LOUVAIN), in both we collect the same number of communities. This means that in students' environment we have a strong clustering in function of place frequency.

5 Conclusions and next steps

After community's discovery, in function of students' behaviors analysis, we want to correlate them with their performance. The first and simplest method we will use to calculate the performance is to combine exam results with their importance and their timeliness (e.g. it penalizes exams that are late on schedule):

$$p(x) = \sum_{n=1}^{k} \left(R_n * Cr_n * \frac{Cy_n}{Sy(x)} \right)$$

k = number of exam successes in session;
R_{i_m} exam result
Cr_i = exam credits
Cy_i = Year of exam course;
$Sy(x)$ = Year of student course

Finally, based on this score and on cluster membership descripted above, we will profile the students. This process will be done studying how to optimize the accuracy of profiling. Defining such accuracy in a well-founded way is part of the challenge.

References

[1] L.Kovanen, J.Saram and K.Kaski, "Reciprocity of mobile phone calls.," *JDySES,* vol. 2, no. 2, pp. 138-151, 2011.

[2] D.J.Watts and S.H.Strogatz, "Collective dynamics of small-world networks," *Nature,* vol. 393, no. 6684, pp. 440-442, 1998.

[3] K.Kianmehr and R.Alhajj, "Calling communities analysis and identification using machine learning techniques," *Expert Systems with Applications,* vol. 36, no. 3, pp. 6218-6226, 2009.

[4] L.Sweeney, "k-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* vol. 10, no. 5, pp. 557-570, 2012.

[5] D.Gaševi´c, R.Janzen and A.Zouaq, "Choose your classmates, your GPA is at stake!" The association of cross-class social ties and academic performance," *Am Behav Sci,* vol. 57, no. 10, p. 1460–1479, 2013.

[6] M.P.Vitale, G.C.Porzio and P.Doreian, "Examining the effect of social influence on student performance through network autocorrelation models," *J Appl Stat,* vol. 43, no. 1, p. 115–127, 2016.

[7] PV.Marsden and KE.Campbell, "Measuring tie strength," *Soc Forces,* vol. 63, no. 2, p. 482–501, 1984.

[8] F.Ahmad, N.Hafieza and I.A.A.Aziz, "The Prediction of Students' Academic Performance Using Classification Data Mining Technique," *Applied Mathematical Sciences,* vol. Vol. 9, no. 2015, pp. 6415 - 6426, 2015.

[9] I.Smirnov and S.Thurner, "Formation of homophily in academic performance: students prefer to change their friends rather than performance," *https://arxiv.org/abs/1606.09082,* 2016.

[10] V.Kassarnig, E.Mones, A.Bjerre-Nielsen, P.Sapiezynski, L.D.Dreyer and S.L.Jørgensene, "Academic performance and behavioral patterns," *Epj Data Science,* vol. 7, no. 10, 2018.