Tutorial proposal

Mining (Streams of) Networked Data

Michelangelo Ceci, University of Bari A. Moro
João Gama, LIAAD - INESC Porto, Faculdade de Economia da Universidade do Porto

Title:

Mining (Streams of) Networked Data

Duration:

half-day tutorial (3 hours)

Abstract, including a brief summary of the topic:

Networks have become ubiquitous in several social, economical and scientific fields, ranging from the Internet to social sciences, biology, epidemiology, geography, communication systems, finance and many others. For this reason, in recent years several data mining approaches, specifically designed for tackling predictive and descriptive tasks for network-structured data, have been proposed. The main challenges they have to face with are: i) the inherent dependency of the connected node which introduces some form of autocorrelation ii) the possible dynamic nature of network data which demands for data stream mining algorithms iii) the possible heterogeneity of nodes and edges. In this tutorial we will discuss the different data mining tasks typically considered when mining network data, with particular emphasis on the three main challenges described before. It will conclude with a review of some practical applications of the presented methods in the areas of functional genomics, sensor networks, and social networks.

Motivation, target audience, and interest for the SAC community

The tutorial is of interest to attendees of several SAC tracks, including Data Mining, Data Streams, Networking and Social Network and Media Analysis.

Outline, including a short summary of every section (up to 2 pages). For each topic, indicate the estimated duration, the basic and most relevant literature, and its subtopics

1) Network data (20 minutes)
    a) Introduction to (heterogeneous) networked data
    b) Introduction to the typical data mining tasks (both descriptive and predictive)

2) The problem of Network autocorrelation (50 minutes)
    a) Definition, problems and opportunities
    b) Different forms of network autocorrelation
    c) Connection with spatial autocorrelation and with the semisupervised smoothness assumption
    d) Collective inference

    Relevant literature:
    - Neville J, Jensen D (2007) Relational dependency networks. J Mach Learn Res 8:653–692
    - Stojanova D et al., (2012) Network regression with predictive clustering trees. Data Mining and Knowledge Discovery, vol. 25, p. 378-413
    - Sen P et al. (2008) Collective Classification in Network Data. AI MAGAZINE 29(3)

3) Analysis of Network data Streams (50 minutes)
    a) Challenges in mining networked data,
    b) Online sampling
    c) Evolving centralities and communities
    d) Tracking the dynamics of evolving communities

    Relevant literature:
    - Holme, P., & Saramäki, J. (2012). Temporal networks. Physics reports, 519(3), 97-125.
    - Aggarwal, C., & Subbian, K. (2014). Evolutionary network analysis: A survey.ACM Computing Surveys (CSUR), 47(1), 10

4) Some practical applications in: (60 minutes)
    a) Functional genomics
        - Gene function prediction
    b) Sensor networks
        - Photovoltaic/wind power prediction
    c) Telecommunication networks
        - Mining the evolution of clusters and communities in social networks

Relevant literature:

- Stojanova D et al., (2013) Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction. BMC Bioinformatics, vol. 14, n. 285
- Bessa R et al. (2009) Entropy and Correntropy Against Minimum Square Error in Offline and Online Three-Day Ahead Wind Power Forecasting. IEEE Transactions on Power Systems, Vol. 24, n. 4
- V. Cerqueira, M. Oliveira, J. Gama, A Framework for Analysing Dynamic Communities in Large-scale Social Networks, ICEIS, 2015

Specific goals and objectives

The Tutorial will cover the state of the art in this rapidly growing area of research. The goal is twofold. From one side, we intend to introduce the various forms of autocorrelation in networked data and to present the challenges that they pose to traditional data mining algorithms. To this aim, we will abstract important issues from a number of application domains with various types of linked data. From the other side, we aim to provide the audience with a survey and a comparison of different problems arising when facing with networked data that are produced in the form of a stream. In this respect, we will present the main characteristics in designing streaming algorithms that work with networked data, and illustrative examples of streaming algorithms for a set of data mining tasks, including clustering, classification, prediction, frequent pattern mining and novelty detection. Finally, we will show some case studies which practically show how principles and methods discussed.

Expected background of the audience

We expect that the audience has knowledge of classical data mining tasks (classification, regression, frequent pattern mining) and knowledge of basic concepts of Data Stream processing.

A biographical sketch of the presenter(s) (with full name, address, e-mail, institution, education, publications, and experience in the subject of the tutorial)

Michelangelo Ceci received a "laurea"degree from the University of Bari in 2001. In 2005 he received his Ph.D. degree in Computer Science from the same University. Since 2005 he is an assistant professor at the Dept of Computer Science, University of Bari, Italy. His main research interests are in data mining and machine learning from complex and networked data. He was a

visiting researcher at the University of Bristol (U.K.) and at the JSI (SLO). He has published more than 140 papers in refereed journals and conferences. He is member of the editorial boards of: IJSNM, IJDSN, IJDATS and JAIS. He is Co-Chair of DS 2016 and ECMLPKDD 2017. He has been the program co-chair of five workshops, the organizing committee chair of SEBD 2007, member of the editorial committee of ``Intelligenza Artificiale'' and member of the editorial board of the ECMLPKDD 2014 and 2015 journal tracks.

João Gama received his Licenciado degree from the Fac. of Engineering of the University of Porto, Portugal. In 2000 he received his Ph.D. degree in Computer Science from the Faculty of Sciences of the same University. He joined the Faculty of Economy where he holds the position of Associate Professor, He is also a senior researcher at LIAAD, a group belonging to INESC Porto. He has worked in projects and authored papers in areas related to machine learning, data streams and adaptive learning systems and is a member of the editorial board of international journals in his area of expertise. He served as Co-chair of ECML 2005, DS 2009, ADMA09, IDA 2011, ECMPKDD 2015 and a series of Workshops on KDDS and Knowledge Discovery from Sensor Data with ACM SIGKDD. He is author of a recent book on Knowledge Discovery from Data Streams.

Contacts:
Michelangelo Ceci, Dipartimento di Informatica, University of Bari. Via Orabona, 4, Bari I-70125, Italy. Phone/Fax: (+39) 080 5442285 email: michelangelo.ceci@uniba.it

João Gama, LIAAD-INESC Porto, Rua Dr. Roberto Frias, 378. 4200-378 Porto, Portugal.
Phone: (+351) 222 094 000, Fax : (+351) 222 094 050 email: jgama@fep.up.pt

Audio Visual equipment needed for the presentation

Projector

Teaching materials on the topic by the presenters, such as slides of earlier tutorials or courses (please provide a link only)

Autocorrelation in Spatial/Network data (SALOMON SEMINAR):
http://videolectures.net/solomon_ceci_ssl/

For sensor networks (SUMMER SCHOOL ON DATA SCIENCES FOR BIG DATA):
http://www.ecmlpkdd2015.org/event/2015-07-06/advanced-topics-predicting-structured-outputs
http://www.ecmlpkdd2015.org/summer-school/ss-schedule?login=slides&file=PortoSchool-Dzeroski-Ceci-Handouts.pdf

Network autocorrelation (ECML PKDD 2011 presentation)
http://videolectures.net/ecmlpkdd2011_stojanova_ceci_regression/

Mining networked data (Seminar @ University of Bari):
http://www.di.uniba.it/~ceci/micFiles/courses/bdii/2014-2015/Mining%20Networked%20Data.pdf
(password: bdii1213)

Distributed data stream mining (Keynote talk, ICSS 2014)
http://www.ieee-iscc.org/2014/KeynoteSpeakers/

Big Data Stream Mining (tutorial @ IEEE BigData 2014)
http://cci.drexel.edu/bigdata/bigdata2014/tutorial.htm

Evolving Social Networks: trajectories of communities (Keynote talk, NFMCP 2013)
http://www.di.uniba.it/~nfmcp2013/invited.html