Ann. Data. Sci. (2015) 2(1):1–3
DOI 10.1007/s40745-015-0036-x

Preface

Some Advanced Techniques in Data Science

Yong Shi[1,2,3] · **Yingjie Tian**[1,2]

Published online: 26 May 2015

This issue of 2015, Annals of Data Science (Volume 2, No. 1) presents seven papers from the several areas of data science. They are contributed from 20 authors and the co-authors come from six countries and regions: Australia, Brazil, Iran, Spain, Russia and UK.

The first paper, "Forecasting with Big Data: A Review," by Hossein Hassani1 and Emmanuel Sirimal Silva, presents a comprehensive review on the use of Big Data for forecasting by identifying and reviewing the problems, potential, challenges and most importantly the related applications. Skills, hardware and software, algorithm architecture, statistical significance, the signal to noise ratio and the nature of Big Data itself are identified as the major challenges which are hindering the process of obtaining meaningful forecasts from Big Data. The review finds that at present, the fields of economics, energy and population dynamics have been the major exploiters of Big Data forecasting whilst factor models, Bayesian models and neural networks are the most common tools adopted for forecasting with Big Data. The second paper, "Bayesian Nonparametric Approaches to Abnormality Detection in Video Surveillance," by Vu Nguyen, Dinh Phung, Duc-Son Pham and Svetha Venkatesh, revisits the abnormality detection problem through the lens of Bayesian nonparametric (BNP)

✉ Yong Shi
yshi@ucas.ac.cn

Yingjie Tian
tyj@ucas.ac.cn

1 Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Beijing, China

2 Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing 100190, China

3 College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA

and develop a novel usage of BNP methods for this problem. In data science, anomaly detection is the process of identifying the items, events or observations which do not conform to expected patterns in a dataset. As widely acknowledged in the computer vision community and security management, discovering suspicious events is the key issue for abnormal detection in video surveillance. The important steps in identifying such events include stream data segmentation and hidden patterns discovery. However, the crucial challenge in stream data segmentation and hidden patterns discovery are the number of coherent segments in surveillance stream and the number of traffic patterns are unknown and hard to specify. In particular, this paper employs the infinite hidden Markov model and Bayesian nonparametric factor analysis for stream data segmentation and pattern discovery. In addition, it introduces an interactive system allowing users to inspect and browse suspicious events. The third paper, "Indebted Households Profiling: A Knowledge Discovery from Database Approach," by Rodrigo Arnaldo Scarpel, Alexandros Ladas and Uwe Aickelin, is to employ a knowledge discovery from database process to identify groups of indebted households and describe their profiles using a database collected by the Consumer Credit Counselling Service (CCCS) in the UK. Employing a framework that allows the usage of both categorical and continuous data altogether to find hidden structures in unlabelled data it was established the ideal number of clusters and such clusters were described in order to identify the households who exhibit a high propensity of excessive debt levels.

The forth paper, "Refining a Taxonomy by Using Annotated Suffix Trees and Wikipedia Resources," by Ekaterina Chernyak and Boris Mirkin, presents a step-by-step approach to taxonomy construction. On the first step, the upper layer frame of taxonomy is built manually according to educational materials. On the next steps, the frame is refined at a chosen topic using theWikipedia category tree and articles, both cleaned of noise. This main tool in this is a naturally defined string-to-text relevance score, based on annotated suffix trees. The relevance scoring is used at several tasks: (1) cleaning the Wikipedia tree or page set of noise; (2) allocating Wikipedia categories to taxonomy topics; (3) deciding whether an allocated category should be included as a child to the taxonomy topic, etc. The resulting fragment of taxonomy consists of three parts: the manually set upper layer topic, the adopted part of the Wikipedia category tree and Wikipedia articles as leaves. Every leaf is assigned a set of so-called descriptors; these are phrases explaining aspects of the leaf topic. The method is illustrated by its application to two domains in the area of mathematics: (a) "Probability theory and mathematical statistics", (b) "Numerical mathematics" (both in Russian). The fifth paper, "Estimation of Stress–Strength Reliability for the Generalized Pareto Distribution Based on Progressively Censored Samples," by S. Rezaei, R. Alizadeh Noughabi and S. Nadarajah, deals with the estimation of stress-strength reliability parameter, $R = P(Y < X)$, based on progressively type II censored samples when stress, strength are two independent generalized Pareto random variables. The maximum likelihood estimators, their asymptotic distributions, asymptotic confidence intervals, bootstrap based confidence intervals and Bayes estimators are derived for R. Using Monte Carlo simulations, the MSE, Bayes risk estimators, credible sets and coverage probabilities are computed and compared. The sixth paper, "Event Management for Sensing Enterprises with Decision Support Systems," by Andrés Boza, M. M. E. Alemany, Llanos Cuenca and Angel Ortiz, exposes a decision support system (DSS) to real-time events

and it is possible to start the decision process from scratch in case any unexpected internal and external events take place. An event monitoring and management system should interact with the DSS to manage events that might affect their decisions. It should act as a supra-system to identify when decisions made are still valid or need to be reanalyzed. The traditional configuration of DSS (where they collect internal and external information of the organization and the decision-maker is involved in the decision-making process) should be extended to treat event management using a monitoring and management system, which monitors internal and external information and facilitate the introduction of no monitored events. This monitor and manager systems become more and more necessary due to the incessant incorporation of new technologies that enables the companies to be more context-sensitive. Furthermore, this new and/or more accurate information, which is obtained for the organization, requires a proper management. The last paper, "Novel Approach for Network Traffic Pattern Analysis using Clustering-based Collective Anomaly Detection", by Mohiuddin Ahmed and Abdun Naser Mahmood, formulate the problem of detecting DoS attacks as a collective anomaly which is a pattern in the data when a group of similar data instances behave anomalously with respect to the entire dataset. They propose a framework for collective anomaly detection using a partitional clustering technique to detect anomalies based on an empirical analysis of an attack's characteristics. They validate the approach by comparing its results with those from existing techniques using benchmark datasets.

The Second International Conference on Data Science will be held in Sydney, Australia, August 8–9, 2015 in conjunction with the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 10–13, Hilton, Sydney. There are a number of researchers have submitted the papers to this conference for exchanging their ideas, in which the selected peer-reviewed papers will be published in the journal late.