HACETTEPE ÜNİVERSİTESİ
ENDÜSTRİ MÜHENDİSLİĞİ BÖLÜMÜ

SEMİNER DUYURUSU

Tarih-Saat: 31/03/2014 Pazartesi – 13:30

Yer: Endüstri Mühendisliği Bölümü Konferans Salonu

ENSEMBLE LEARNING STRATEGIES FOR LARGE-SCALE TIME SERIES DATA MINING

Assistant Professor Mustafa Baydoğan
Boğaziçi University

Classical approach to time series analysis accounts for the fact that data points observed over time may have an internal structure (such as autocorrelation, trend or seasonal variation). However, these approaches generally impose assumptions on the data such as stationarity, linearity, normally distributed residuals, etc. Unfortunately, most of the time series data in real-world do not satisfy these assumptions. Also, analysis is problematic when the time series are multivariate and the number of observations is large. During the last decade, the increase in the number of temporal datasets generated in different domains such as medicine, finance, multimedia, etc. has made researchers focus on time series data mining approaches. These approaches mostly use the machine learning techniques to understand and model the time series data. Time series can be modeled with minimum number of assumptions and also parameters using these methods. A time series model is generally referred to as time series representation in this domain.

Considering the strengths and weaknesses of both approaches, we introduce a novel time series representation based on a tree-based ensemble learning strategy. Earlier, many high-level representations such as Fourier transforms, wavelets, piecewise polynomial models etc. have been proposed for time series data mining. However, these representations require many parameters and have problems with generalizability. Our tree-based ensemble learning approach imposes no constraints on the data and has only one parameter. It can handle numerical, categorical and ordinal multivariate inputs/outputs, nonlinear and interactive effects. It is scale invariant and robust to missing values. More importantly, the complexity of the approach is linear to both the length of the series and the number of datasets in a time series database. We illustrate the benefits of our approach on 45 time series classification problems from University of California, Riverside (UCR) time series database. The proposed approach has promising extensions to forecasting, clustering, anomaly detection etc. Also, an R package named 'LPStimeSeries' is made available at The Comprehensive R Archive Network (http://cran.r-project.org/web/packages/LPStimeSeries/index.html)

Key words: time series, pattern discovery, regression tree, ensemble learning

Bio: Mustafa Gökçe Baydoğan is an assistant professor in Department of Industrial Engineering at Boğaziçi University, Istanbul, Turkey. Before joining Boğaziçi University, he worked as a postdoctoral research assistant in the Security and Defense Systems Initiative at Arizona State University (ASU) between 2012-2013. He received his Ph.D. degree in Industrial Engineering from ASU in 2012. His B.S. and M.S. degrees are in Industrial Engineering both from Department of Industrial Engineering at Middle East Technical University, Ankara, Turkey in 2006 and 2008 respectively. His current research interests focus on statistical learning, with applications in temporal data mining (time series and images) and data mining for massive, multivariate data sets.

Details about him and his work can be reached through: www.mustafabaydogan.com.