Lokacija: Fakultet Organizacionih Nauka (Jove Ilica 154)

Termin/Sala: 13:00h, 15. Juni 2009, Novi Amfiteatar - 2

Dynamic Clustering-Based Estimation of Missing Values in Mixed Type Data

Zoran Obradovic

Director, Center for Information Science and Technology, Temple University, Philadelphia, USA

ABSTRACT: The appropriate choice of a method for imputation of missing data becomes especially important when the fraction of missing values is large and the data are of mixed type. The proposed dynamic clustering imputation (DCI) algorithm relies on similarity information from shared neighbors, where mixed type variables are considered together. When evaluated on a public social science dataset of 46,043 mixed type instances with up to 33% missing values, DCI resulted in more than 20% improved imputation accuracy over Multiple Imputation, Predictive Mean Matching, Linear and Multilevel Regression, and Mean Mode Replacement methods. Data imputed by 6 methods were used for test of NB-Tree, Random Subset Selection and Neural Network-based classification models. In our experiments classification accuracy obtained using DCI-preprocessed data was a lot better than when relying on alternative imputation methods for data preprocessing.

The reported results were obtained through a collaboration with Vadim Ayuyev, Joseph Jupin and Philip W. Harris funded by research grant No. 2006-IJ-CX-0022 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice.

SPEAKER: Zoran Obradovic, professor of computer and information sciences and the director of the Center for Information Science and Technology at Temple University in Philadelphia is an internationally recognized leader in data mining and bioinformatics. He joined Temple University in 2000 from Washington State University, where he was named Researcher of the Year by the College of Engineering and Architecture. At Temple University in 2008 he received College of Science and Technology Faculty Research Excellence Award and in 2009 the overall Temple University Faculty Research Award. His research focuses on improving predictive modeling and decision support through data-driven discovery and modeling of hidden patterns in large data sets. Obradovic has published more than 200 articles addressing data mining challenges in health informatics, the social sciences, environmental management and other domains. His group's pioneering research on the prediction and functional analysis of intrinsically disordered regions in proteins has provided new insight into how protein structure establishes function and the program his team developed was the best rated predictor of intrinsic disorder at three consecutive international competitions organized by protein structure prediction assessment community (CASP 5-7). Obradovic was the program chair at five, track chair at seven and program committee member at about 40 international conferences on data mining. He currently serves as an editorial board member at seven journals.