Matematički Fakultet, Univerzitet u Beogradu,
Studentski trg 16, sala 718/IV sprat
termin: 12:00h, 2. septembar 2011.

Learning from Inconsistent and Unreliable Annotators by a Gaussian Mixture Model and Bayesian Information Criterion

Zoran Obradović

Director, Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, USA

ABSTRACT: Supervised learning from multiple annotators is an increasingly important problem in machine leaning and data mining. In this lecture we will describe our recently developed probabilistic approach to this problem when annotators are not only unreliable, but also have varying performance depending on the data. The proposed approach uses a Gaussian mixture model and Bayesian information criterion to find the fittest model to approximate the distribution of the instances. Then the maximum a posterior estimation of the hidden true labels and the maximum-likelihood estimation of quality of multiple annotators are provided alternately. Experiments on emotional speech classification and CASP 9 protein disorder prediction tasks show performance improvement of the proposed approach as compared to the majority voting baseline and a previous data independent approach. Moreover, our approach also provides more accurate estimates of individual annotators performance for each Gaussian component, thus paving the way for understanding the behaviors of each annotator.

Presented results are obtained in collaboration with my Ph.D. student Ping Zhang at Temple University. Articles providing additional details are in press and will appear at the *Proteome Science* journal and at the *Proceedings of the 2011 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*.

SPEAKER INFORMATION: Zoran Obradovic, professor of Computer and Information Sciences and the director of the Center for Data Analytics and Biomedical Informatics at Temple University in Philadelphia is an internationally recognized leader in data mining and bioinformatics. He has published about 230 articles addressing data mining challenges in health informatics, climate and ecological management, the social sciences, and other domains. Obradovic was the program chair at six, track chair at seven and program committee member at about 40 international conferences on data mining. He is an editorial board member at seven journals and is the executive editor at the journal on Statistical Analysis and Data Mining which is the official publication of the American Statistical Association (ASA).