

Learning from Imbalanced Data: An Overview and the DataBoost-IM Algorithm

Learning from imbalanced data sets, where the number of examples of one (majority) class is much higher than the others, presents an important challenge to the machine learning community. Many real world applications involve learning from imbalanced sets, such as fraud detection, oil spill detection and text classification. Traditional machine learning algorithms may be biased towards the majority class, thus producing poor predictive accuracy over the minority class.

In this talk we overview of techniques that aim to address the class imbalance problem. We also describe our method that combines boosting, an ensemble-based learning algorithm, with data generation to improve the predictive power of classifiers against imbalanced data sets consisting of two classes.

Short Biography

Herna L Viktor is an associate professor at the School of IT and Engineering (SITE), University of Ottawa, Canada. She is a member of the Text Analysis and Machine Learning (TAMALE) group and the leader of the Intelligent Decision Support and Data Analysis Lab (IDeAL) at SITE. Her research focuses on the development of new methodologies for the management and data mining of large-scale object-relational databases and data warehouses. The end results of her research have been applied within the Anthropometry, Health Care and Bioinformatics domains. She holds a Ph. D. in Computer Science from the University of Stellenbosch, which she received in 1999, has published more than 120 international journal and conference articles and is on a number of international programme committees. Her research is sponsored by the Canadian National Science and Engineering Research Council (NSERC), the Canada Foundation for Innovation (CFI) and the Ontario Innovation Trust (OIT).