Jean-Charles Lamirel – Pascal Cuxac – Vincent Lemaire
LORIA (France) – CNRS inist (France) – Orange Labs (France)

Plenary talk:

Zhi-Hua Zhou

http://cs.nju.edu.cn/zhouzh/
Nanjing University Nanjing, China

Title "Incrementally Optimizing AUC"

Zhi-Hua Zhou is a Cheung Kong professor at Nanjing University. His research interests mainly include machine learning, data mining, pattern recognition and multimedia information retrieval. He has published more than 100 papers, authored the book "Ensemble Methods: Foundations and Algorithms" (2012), and holds 12 patents. According to GoogleScholar, his publications have received more than 11,000 citations, with an h-index 54.
He is the recipient of the IEEE CIS Outstanding Early Career Award, Fok Ying Tung Young Professorship Award, Microsoft Young Professorship Award, and various awards including nine international journal/conference paper or competition awards.
He serves/ed as Executive Editor-in-Chief of "Frontiers of Computer Science", Associate Editor-in-Chief of "Chinese Science Bulletin", Associate Editor or Editorial Boards member of "ACM TIST", "IEEE TKDE" and many other journals. He is the Founder of ACML, and Steering Committee member of PAKDD and PRICAI. He served as Area Chair or PC member for almost all top conferences in his areas. He is the Chair of the AI&PR Technical Committee of the China Computer Federation, Chair of the Machine Learning Technical Committee of the China Association of AI, Vice Chair of the Data Mining Technical Committee of the IEEE Computational Intelligence Society, and Chair of the IEEE Computer Society Nanjing Chapter. He is a Fellow of the IAPR and Fellow of the IEEE.

Session 1: Classifier combination in dynamic context

Incremental Ensemble Classifier Addressing Dynamic, Evolving, and Non-Stationary Streams

Brandon Parker[1], Latifur Khan[1], Albert Bifet[2].

[1]University of Texas at Dallas, USA

[2]Huawei, Hong Kong

Classification of data points in a data stream is a fundamentally different set of challenges than data mining on static data. While streaming data is often placed into the context of "Big Data" wherein one-pass algorithms are used, true data streams offer additional hurdles due to their dynamic, evolving, and non-stationary nature. During the stream, the available labels (or concepts) often change, and a concept's definition in the feature space can also evolve (or drift) over time. The core issue is that the hidden generative function of the data is not a constant function, but rather the background data distribution also evolves over time. This is known as a non-stationary distribution. In this paper, we describe a new approach to using ensembles for stream classification. While the core method is straightforward, it is specifically designed to adapt quickly with very little overhead to the dynamic and evolving nature of data streams generated from non-stationary functions. Our method, M3, is based on a weighted majority ensemble of heterogeneous model types where model weights are updated online using Reinforcement Learning techniques. We compare our method with current leading algorithms as implemented in the Massive Online Analysis (MOA) framework using UCI benchmark and synthetic stream generator data sets, and find that our method shows particularly strong gain over the baseline method when ground truth is of limited availability to the classifiers.

Merging Classifiers of Different Classification Approaches

Antonina Danylenko[1], Welf Löwe[1]

[1]Linnaeus University, Sweden

Classification approaches, e.g. decision trees or Naive Bayesian classifiers, are often tightly coupled to learning strategies, special data structures, the type of information captured, and to how common problems, e.g. overfitting, are addressed. This prevents a simple combination of classifiers of different approaches learned over different data sets. Many different methods of combining classification models have been proposed. However, most of them are based on a combination of the actual result of classification rather than producing a new, possibly more accurate, classifier capturing the combined classification information. In this paper we propose a new general approach to combining different classification models based on a concept of Decision Algebra which provides a unified formalization of classification approaches as higher order decision functions. It defines a general combining operation, referred to as merge operation, abstracting from implementation details of different classifiers. We show that the combination of a series of probably accurate decision functions (regardless of the actual implementation) is even more accurate. This can be exploited, e.g., for distributed learning and for efficient general online learning. We support our results by combining a series of decision graphs and Naive Bayesian classifiers learned from random samples of the data sets. The result shows that on each step the accuracy of the combined classifier increases, with a total accuracy growth of up to 17%.

Coffee break

Session 2: Drift and anomaly detection

Drift Detection for Multi-label Data Streams Based on Label Grouping and Entropy

Shi Zhongwei[1], Wen Yimin[1], Feng Chao[1], Zhao Hai[2].

[1]University of Electronic Technology, China

[2]Shanghai Jiao Tong University, China

Algorithms that detect concept drift are important for real-world applications and many of them involve data which can be considered as multi-label data streams. Effective drift detection methods should be able to consider the unique properties of multi-label stream data, such as label dependence and multiple types of concept drift. To deal with these challenges, we present an efficient and effective method of detecting concept drift based on label grouping and entropy for multi-label data streams. It employs the potential and available methods to group the set of labels into different subsets and adopts a multi-label version of entropy to measure the distribution of multi-label data. Then it detects concept drift by comparing the entropies of the older and more current multi-label data. Our particular analysis and discussion are based on three synthetic datasets with different types of concept drift and the proposed method is also tested with real-world datasets. The experimental results show a better performance for detecting drift of the proposed method compared with the baseline methods.

Efficient Anomaly Detection by Isolation Using Nearest Neighbour Ensemble

Tharindu Bandaragoda[1], Kai Ming Ting[2], David Albrecht[1], Fei Tony Liu[1], Jonathan Wells[1].

[1]Monash University, Australia

[2]Federation University, Australia

This paper presents iNNE (isolation using Nearest Neighbour Ensemble), an efficient nearest neighbour-based anomaly detection method by isolation. iNNE runs significantly faster than existing nearest neighbour-based methods such as Local Outlier Factor, especially in data sets having thousands of dimensions or millions of instances. This is because the proposed method has linear time complexity and constant space complexity. Compared with the existing tree-based isolation method iForest, the proposed isolation method overcomes three weaknesses of iForest that we have identified, i.e., its inability to detect local anomalies, anomalies with a low number of relevant attributes, and anomalies that are surrounded by normal instances.

Session 3: Roundtable – discussions