ECML 2007 PKDD
WARSAW POLAND

The 18th European Conference on Machine Learning
and
the 11th European Conference on Principles and Practice
of Knowledge Discovery in Databases

Proceedings of the International Workshop on Knowledge Discovery from Ubiquitous Data Streams

IWKDUDS 2007

September 17, 2007

Warsaw, Poland

Editors:
João Gama
LIAAD - INESC Porto L.A. & University of Porto, Portugal
Mohamed Medhat Gaber
Tasmanian ICT Centre, CSIRO ICT Centre, Australia
Jesús S. Aguilar-Ruiz
School of Engineering, Pablo de Olavide University, Seville, Spain

Preface

We are glad to have this year's international workshop on knowledge discovery from ubiquitous data streams at ECML/PKDD 2007. We have a strong workshop program with 12 papers, an invited talk and a tutorial.

Ting, Theodorou and Schaal have introduced a modified Kalman Filter that can perform real-time outlier detection. Spinosa, Carvalho, and Gama have extended OLINDDA (OnLIne Novelty and Drift Detection Algorithm) to multiple-class classification problems. Lei, Tang, Iglesias, Mukherjee, and Mohanty have presented a simlilarity-driven clustering approach to address the scalalbility probelms in large datasets with an application to Gravitational-Wave Astronomy Data. Yoshida and Hruschka Jr. have decribed and evaluated experimentally a Quasi-Incremental Bayesian Classifier that could be used in dynamic systems like sensor networks. Küçük, Inan, Boyrazoglu, Buhan, Salor, Çadirci, and Ermis have presented a data stream architecture for electrical power quality (PQStream). Phung, Gaber and Roehm have extended their ERA-Cluster clusiting algorithm to work in a distributed mode in wireless sensor networks. Karnstedt, Franke and Gaber have introduced and described mathematically a model for quality guaranteed resource-aware stream mining. Landwehr, Gutmann, Thon, Philipose, and Raedt have described a ubiquitous computing application to recognize human activities from sensory data. Last and Saveliev have enhanced the Information Network (IN) classification algorithm by preserving the model qulaity while reducing the coputational cost. Rodrigues and Gama have extended their clustering technique Online Divisive-Agglomerative Clustering (ODAC) using semi-fuzzy approach. Haghighi, Gaber, Krishnaswamy, Zaslavsky, and Loke have introduced an architecture for context-aware adaptive data stream mining. Finally, An and Park have introduced an efficient secure XML query processing method.

We hope that this proceedings will form an important and valuable addition to your library. Finally, we thank all the authors for their significant contributions to the workshop.

August 2007

João Gama
Mohamed Medhat Gaber
Jesús S. Aguilar-Ruiz

Workshop Organization

Workshop Chairs

João Gama (LIAAD - INESC Porto L.A. & University of Porto, Portugal)
Mohamed Medhat Gaber (Tasmanian ICT Centre, CSIRO ICT Centre, Australia)
Jesús S. Aguilar-Ruiz (School of Engineering, Pablo de Olavide University, Spain)

Publicity Chair

Pedro Pereira Rodrigues (LIAAD - INESC Porto L.A. & University of Porto, Portugal)

ECML/PKDD Workshop Chair

Marzena Kryszkiewicz (Warsaw University of Technology)

Workshop Program Committee

Andreas Hotho
André Carvalho
Antoine Cornuejols
Bernhard Seeger
Elaine Sousa
Eduardo Spinosa
Francisco Ferrer-Troyano
Auroop Ganguly
Geoff Holmes
Georges Hebrail
Hillol Kargupta
João Gama
Jesús Aguilar-Ruiz
Josep Roure
Mark Hall
Mark Last
Miroslav Kubat
Mohamed Medhat Gaber
Olufemi Omitaomu
Pedro Pereira Rodrigues
Philip S. Yu
Ralf Klinkenberg
Rasmus Pedersen
Ricard Gavalda
Sean Wang
Takashi Washio
Raju Vatsavai
Ying Yang

Table of Contents

Learning an Outlier-Robust Kalman Filter

Jo-Anne Ting[1], Evangelos Theodorou[1], and Stefan Schaal[1,2]

[1] University of Southern California
Los Angeles, CA 90089, USA
joanneti,etheodor@usc.edu
[2] ATR Computational Neuroscience Laboratories
Kyoto, Japan
sschaal@usc.edu

Abstract. In this paper, we introduce a modified Kalman filter that performs robust, real-time outlier detection, without the need for manual parameter tuning by the user. Systems that rely on high quality sensory data (for instance, robotic systems) can be sensitive to data containing outliers. The standard Kalman filter is not robust to outliers, and other variations of the Kalman filter have been proposed to overcome this issue. However, these methods may require manual parameter tuning, use of heuristics or complicated parameter estimation procedures. Our Kalman filter uses a weighted least squares-like approach by introducing weights for each data sample. A data sample with a smaller weight has a weaker contribution when estimating the current time step's state. Using an incremental variational Expectation-Maximization framework, we learn the weights and system dynamics. We evaluate our Kalman filter algorithm on data from a robotic dog.

Full article in PDF

Learning novel concepts: beyond one-class classification with OLINDDA

Eduardo J. Spinosa[1], André Ponce de Leon F. de Carvalho[1], and João Gama[2]

[1] University of São Paulo (USP), Institute of Mathematical and Computer Sciences
(ICMC), Caixa Postal 668, 13560-970, São Carlos, SP, Brazil
ejspin@icmc.usp.br, andre@icmc.usp.br
www.icmc.usp.br/~ejspin, www.icmc.usp.br/~andre

[2] University of Porto (UP), Artificial Intelligence and Computer Science Laboratory (LIACC),
Rua Campo Alegre, 823, 4150, Porto, Portugal
jgama@liacc.up.pt
www.liacc.up.pt/~jgama

Abstract. OLINDDA (OnLIne Novelty and Drift Detection Algorithm) addresses the problem of novelty detection in an online continuous learning scenario as an extension to a single-class classification problem. This paper presents its current version, that evolved toward the discovery of new concepts initially as emerging clusters and further as cohesive sets of clusters. New strategies for validation and merging of clusters as well as for dynamically adapting the number of clusters are discussed and experimentally evaluated.

Key words: Novelty detection, Unsupervised learning, Clustering, K-Means

Full article in PDF

S-means: Similarity Driven Clustering and Its application in Gravitational-Wave Astronomy Data Mining

Hansheng Lei[1], Lappoon R. Tang[1], Juan R. Iglesias[1]
Soma Mukherjee[2], and Soumya Mohanty[2]

[1] Computer Science Department
[2] The Center for Gravitational Wave Astronomy
The University of Texas at Brownsville
Brownsville TX 78520, USA
hansheng.lei@utb.edu

Abstract. Clustering is to classify unlabeled data into groups. It has been well-researched for decades in many disciplines. Clustering in massive amount of astronomical data generated by multi-sensor networks has become an emerging new challenge; assumptions in many existing clustering algorithms are often violated in these domains. For example, K means implicitly assumes that underlying distribution of data is Gaussian. Such an assumption is not necessarily observed in astronomical data. Another problem is the determination of K, which is hard to decide when prior knowledge is lacking. While there has been work done on discovering the proper value for K given only the data, most existing works, such as X-means, G-means and PG-means, assume that the model is a mixture of Gaussians in one way or another. In this paper, we present a similarity-driven clustering approach for tackling large scale clustering problem. A similarity threshold T is used to constrain the search space of possible clustering models such that only those satisfying the threshold are accepted. This forces the search to: 1) explicitly avoid getting stuck in local minima, and hence the quality of models learned has a meaningful lower bound, and 2) discover a proper value for K as new clusters have to be formed if merging them into existing ones will violate the constraint given by the threshold. Experimental results on the UCI KDD archive and realistic simulated data generated for the Laser Interferometer Gravitational Wave Observatory (LIGO) suggest that such an approach is promising.

Full article in PDF

Quasi-Incremental Bayesian Classifier

Murilo Lacerda Yoshida[1] and Estevam R. Hruschka Jr.[1]

DC / UFSCar, Universidade Federal de São Carlos, São Carlos, Brazil
{murilo_yoshida, estevam}@dc.ufscar.br

Abstract. This paper describes and empirically evaluates a Quasi-Incremental Bayesian Classifier (QBC) designed to be used when a classification task must be performed in dynamic systems such as sensor networks, which are continuously receiving new piece of information to be stored in huge databases. Therefore, the knowledge that needs to be extracted from these databases is continuously evolving and the learning process may need to go on almost indefinitely. The induction proposed by QBC is performed in two steps; in the first one a traditional Bayesian Network (BN) induction algorithm is performed using an initial amount of data. As far as new data is available, only the numerical parameters of the classifier are updated. The conducted experiments showed that QBC tends to maintain the average correct classification rates obtained with non-incremental classifiers while decreasing the time needed to induce the classifier.

Key words: Bayesian Networks, Bayesian Classifiers, Incremental Learning.

Full article in PDF

PQStream: A Data Stream Architecture for Electrical Power Quality

Dilek Küçük[1], Tolga İnan[1], Burak Boyrazoğlu[1,2], Serkan Buhan[1,3]
Özgül Salor[1], Işık Çadırcı[1,3], and Muammer Ermiş[2]

[1] TÜBİTAK – Uzay, Power Quality Group, Ankara – Turkey
{dilek.kucuk, tolga.inan, burak.boyrazoglu, serkan.buhan,
ozgul.salor}@uzay.tubitak.gov.tr
[2] METU, Electrical and Electronics Eng. Dept., Ankara – Turkey
ermis@metu.edu.tr
[3] Hacettepe University, Electrical and Electronics Eng. Dept., Ankara – Turkey
cadirci@ee.hacettepe.edu.tr

Abstract. In this paper, a data stream architecture is presented for electrical power quality (PQ) which is called PQStream. PQStream is developed to process and manage time-evolving data coming from the country-wide mobile measurements of electrical PQ parameters of the Turkish Electricity Transmission System. It is a full-fledged system with a data measurement module which carries out processing of continuous PQ data, a stream database which stores the output of the measurement module, and finally a Graphical User Interface for retrospective analysis of the PQ data stored in the stream database. The presented model is deployed and is available to PQ experts, academicians and researchers of the area. As further studies, data mining methods such as classification and clustering algorithms will be applied in order to deduce useful PQ information from this database of PQ data.

Key words: Data Streams, Data Stream Applications, Electrical Power Quality.

Full article in PDF

Resource-aware Distributed Online Data Mining for Wireless Sensor Networks

Nhan Duc Phung[1], Mohamed Medhat Gaber[2], and Uwe Roehm[1]

[1] University of Sydney, School of Information Technologies
SIT Building J12, NSW 2006, Australia
{dphu9727,roehm}@it.usyd.edu.au
[2] CSIRO ICT Centre, Tasmania, Hobart, TAS 7001
Mohamed.Gaber@csiro.au

Abstract. Online data mining in wireless sensor networks is concerned with the problem of extracting knowledge from a large continuous amount of data streams with an in-network processing mode. Unlike other types of networks, the limited computational resources require the mining algorithms to be highly efficient and compact. We propose a distributed resource-aware online data mining framework for wireless sensor networks which can be used to enable existing mining techniques to be applied to sensor network environments. We have applied the framework to develop and implement a distributed resource adaptive online clustering algorithm on the novel Sun MicrosystemTM Small Programmable Object Technology Sun SPOT platform. We have evaluated the performance of the algorithm on the actual sensor nodes. Experimental results show that the clustering algorithm can improve significantly in resource utilization while maintaining acceptable accuracy level.

Key words: distributed clustering, resource adaptivity, data mining, sensor networks

Full article in PDF

A Model for Quality Guaranteed Resource-Aware Stream Mining

Marcel Karnstedt[1], Conny Franke[2], and Mohamed Medhat Gaber[3]

[1] Technische Universität Ilmenau, Ilmenau, Germany
[2] University of California at Davis, Davis, CA, USA
[3] Tasmanian ICT Centre, CSIRO ICT Centre, Australia

Abstract. Data streams are produced continuously at a high speed. Most data stream mining techniques address this challenge by using adaptation and approximation techniques. Adapting to available resources has been addressed recently. Although these techniques ensure the continuity of the data mining process under resource limitation, the quality of the output is still an open issue. In this paper, we propose a generic model that guarantees the quality of the output while maintaining efficient resource consumption. The model works on estimating the quality of the output given the available resources. Only a subset of these resources will be used that guarantees the minimum quality loss. The model is generalized for any data stream mining technique.

Full article in PDF

Relational Transformation-based Tagging for Human Activity Recognition

Niels Landwehr[1], Bernd Gutmann[1], Ingo Thon[1]
Matthai Philipose[2], and Luc De Raedt[1]

[1] Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200 A, B-3001 Heverlee, Belgium
firstname.lastname@cs.kuleuven.be
[2] Intel Research Seattle
1100 NE 45th Street
Seattle, WA 98105, USA
matthai.philipose@intel.com

Abstract. The ability to recognize human activities from sensory information is essential for developing the next generation of smart devices. Many human activity recognition tasks are from a machine learning perspective quite similar to tagging tasks in natural language processing. Motivated by this similarity, we develop a relational transformation-based tagging system based on inductive logic programming principles, which is able to cope with expressive relational representations as well as a background theory. The approach is experimentally evaluated on two activity recognition tasks and compared to Hidden Markov Models, one of the most popular and successful approaches for tagging.

Full article in PDF

Enhanced Anytime Algorithm for Induction of Oblivious Decision Trees

Mark Last[1] and Albina Saveliev[1]

Department of Information Systems Engineering
Ben-Gurion University of the Negev
POB 653, Beer-Sheva, 84105 Israel
{mlast, albinabu}@bgu.ac.il

Abstract. Real-time data mining of high-speed and non-stationary data streams has a large potential in such fields as efficient operation of machinery and vehicles, wireless sensor networks, urban traffic control, stock data analysis etc.. These domains are characterized by a great volume of noisy, uncertain data, and restricted amount of resources (mainly computational time). Anytime algorithms offer a tradeoff between solution quality and computation time, which has proved useful in applying artificial intelligence techniques to time-critical problems. In this paper we are presenting a new, enhanced version of an anytime algorithm for constructing a classification model called Information Network (IN). The algorithm improvement is aimed at reducing its computational cost while preserving the same level of model quality. The quality of the induced model is evaluated by its classification accuracy using the standard 10-fold cross validation. The improvement in the algorithm anytime performance is demonstrated on several benchmark data streams.

Key words: anytime algorithms, classification, information theory, Information Network algorithm, classification accuracy, computation cost

Full article in PDF

A Semi-Fuzzy Approach for Online Divisive-Agglomerative Clustering

Pedro Pereira Rodrigues[1,2] and João Gama[1,3]

[1] LIAAD - INESC Porto L.A.
[2] Faculty of Sciences of the University of Porto
[3] Faculty of Economics of the University of Porto
Rua de Ceuta, 118 - 6 andar, 4050-190 Porto, Portugal
pprodrigues@fc.up.pt jgama@fep.up.pt

Abstract. The Online Divisive-Agglomerative Clustering (ODAC) is an incremental approach for clustering streaming time series using a hierarchical procedure over time. It constructs a tree-like hierarchy of clusters of streams, using a top-down strategy based on the correlation between streams. The system also possesses an agglomerative phase to enhance a dynamic behavior capable of structural change detection. However, the split decision used in the algorithm focus on the crisp boundary between two groups, which implies a high risk since it has to decide based on only a small subset of the entire data. In this work we propose a semi-fuzzy approach to the assignment of variables to newly created clusters, for a better trade-off between validity and performance. Experimental work supports the benefits of our approach.

Key words: fuzzy clustering, streaming time series, hierarchical models.

Full article in PDF

An Architecture for Context-Aware Adaptive Data Stream Mining

Pari Delir Haghighi, Mohamed Medhat Gaber, Shonali Krishnaswamy
Arkady Zaslavsky, and Seng Loke

[1] Center for Distributed Systems and Software Engineering
Monash University, Australia
{pari.delirhaghighi, shonali.krishnaswamy, Arkady Zaslavsky}@infotech.monash.edu.au
[2] CSIRO ICT Center, Australia
Mohamed.gaber@csiro.au
[3] Department of Computer Science and Computer Engineering
La Trobe University, Australia
s.loke@latrobe.edu.au

Abstract. In resource-constrained devices, adaptation of data stream processing to variations of data rates, availability of resources and environment changes is crucial for consistency and continuity of running applications. Context-aware adaptation, as a new dimension of research in data stream mining, enhances and optimizes distributed data stream processing tasks. Context-awareness is one of the key aspects of ubiquitous computing as applicationsÇ successful operations rely on detecting changes and adjusting accordingly. This paper presents a general architecture for context-aware adaptive mining of data streams that aims to dynamically and autonomously adjust data stream mining parameters according to changes in context and resource availability in distributed and heterogeneous computing environments.

Full article in PDF

Efficient Secure Query Processing in XML Data Stream

Dong-Chan An and Seog Park

Department of Computer Science & Engineering, Sogang University
C.P.O. Box 1142, Seoul Korea 100-611
{channy, spark}@sogang.ac.kr

Abstract. As various users and applications require the distribution and sharing of information in XML documents, the need for an efficient secure access of XML data in a ubiquitous data stream environment has become very important. In this paper, we propose an efficient secure XML query processing method to solve the two problems by using role-based prime number labeling and XML fragmentation. A medical records XML document has the characteristic of an infinite addition in width rather than in depth because of the increment of patients. But a role-based prime number labeling method can fully manage the size of documents that increases to infinity and can minimize the maintenance cost caused by dynamic changes. Experimental evaluation clearly demonstrates that our approach is efficient and secure.

Full article in PDF

Author Index