AI Communications 29 (2016) 621–625
DOI 10.3233/AIC-160718
IOS Press

Guest Editorial

Special Issue on Environmental Data Mining

Karina Gibert
CCIA'2013 PC-Chair, Guest Editor, Knowledge Engineering and Machine Learning Group, Department Statistics and Operation Research, Universitat Politècnica de Catalunya-BarcelonaTech, Barcelona, Spain
E-mail: karina.gibert@upc.edu

Data Mining is the discipline for *non trivial identifying of valid, novel, potentially useful, ultimately understandable patterns in data* [3] and provides the opportunity to extract relevant decisional knowledge from data bases in any application field. In particular it can contribute to a better understanding of Environmental Sciences. Environmental Sciences is a wide research field with a number of open problems that require attention. The technological development occurred in the last years has significantly increased the availability of environmental data to be exploited for better addressing current challenges in the area.

My interest in data mining comes from the 90s, when I was a PhD student. From then, part of my research activity has been focused on disseminating data mining in different areas of application, as well as to contribute to these areas by applying data mining to some challenging real problems. Among other activities, I'm the chair of the *Data Mining Techniques for Environmental Sciences (DMTES)* workshop, which we organize every two years in the frame of the *Environmental Modelling and Software Society* biannual meetings. The DMTES series started in 2002, with the aim of becoming a multidisciplinar discussion forum for the Data Mining and Environmental Sciences communities. From then onwards, every two years the DMTES has been celebrated all along the world (Vermont (USA), Barcelona (Spain), Otawa (Canada), Leizig (Germany), San Diego (USA), Toulouse (France)), providing a valuable opportunity for a close contact between the data mining community and the Environmental Sciences community, and discussions arising in the workshop raise current challenges in both areas and synergic opportunities. From these meetings and other collaborations in environmental applications we learned that potentials of Data Mining in Environmental Sciences are enormous, and they are still poorly exploited by many reasons, in part due to a lack of knowledge of these potentials from the environmental side, and also to a lack of knowledge of the current environmental challenges from the Data Mining side.

This special issue launches with the spirit of giving a clue on how data mining can contribute to a better comprehension and management of the environment as well as the kind of environmental challenges that can be addressed from a data mining perspective.

The issue contains a careful selection of 8 papers organized around different topics, with a good coverage of the different research topics currently active in Environmental Data Mining, from more theoretical to more applied, from more specific applications, to more global ones, and covering a perspective of different environmental systems and data mining methods. Two of the papers and are surveys in challenging topics for environmental data mining, one from the data mining methodologies perspective, the other from the environmental sciences perspective. The first survey [6] is about preprocessing data, which is the very first step in any data mining project. The second one [7] is about the use of data mining methods in water management decision support, which is the very last step of the data mining process.

From the point of view of the environmental systems targeted in the different papers, the issue shows research on climate change, ecological systems, ichtiology, water quality, water management and land use. The papers address a variety of environmental problems like understanding the effects of water quality in freshwater macroinvertebrates occurrence, fish species

identification, water systems monitoring and potencial sensor faults, surface water quality prediction, heatwave identification, urban dynamic analysis.

From the point of view of the data mined in the different papers, the issue shows works analyzing classical data matrices, but in almost all the works built after some previous work. Data sharing is required to build data matrices merging information from international data bases. Also works using open temporal data, images, sensor data, topographic data bases are presented. Some works use off-line data, other work with online data. In all the works the preprocessing step is shown as a relevant step of the process, sometimes for merging several pre-existent databases, sometimes to apply feature extraction techniques for transforming either images or topographic objects into a classical data matrix, sometimes to validate data quality, sometimes to multigranular treatment of temporal data, somethimes to label geolocalized objects, sometimes to deal with unbalanced data. From the point of view of the data mining models used, some works use classification methods, times series forecasting, sequential pattern mining, clustering, spatio-temporal clustering, semantic distances, classification and regression trees or visualization in a wide overview of different kind of datamining methods useful in environmental systems analysis. Only a couple of works provide explicit reference to the post-processing of data mining results towards an understandable results description, and only one work is tackling an explicit step of new knowledge production, but these topics are really relevant in environmental data mining. Finally three of the works mention in an explicit and direct way the global goal of supporting further decision making processess in the fields of management of natural resources and territories (landscape) and urban planning, allocation of river restoration resources and preservation of climate change oriented policy making.

Thus, the works in this issue show how the set of steps composing the whole data mining process can be addressed in a variety of environmental systems, with different environmental problems and different types of data, giving a good overview of the current activity in the field.

The paper *A survey on pre-processing techniques: relevant issues in the context of environmental data mining* stresses the importance of the very first step of a data mining process that too often receive poor attention. Indeed, the pre-processing step is critical for a proper data preparation that guarantees a correct analysis, even if few literature is devoted to it. The paper identifies the main aspects to be considered in preprocessing in the context of environmental data mining and proposes a methodological framework to perform the preprocessing in a systematic way. It also provides an overview of the state of the art in this topics, identifies current challenges in the area, and provides guidelines for both data miners and end-users.

The paper *Analysing the effects of water quality on the occurrence of freshwater macroinvertebrate taxa among tropical river basins from different continents* contributes to a deeper understanding of the specificity and universality of ecological models. Therefore, the responses of macroinvertebrates towards pollution among different countries/regions of similar climatic conditions were analyzed by means of classification trees. Such quantitative relationships give a more reliable global and local understanding of ecosystem responses towards pollution and climate change for instance. The main findings of the authors were that responses of macroinvertebrates towards pollution were different among countries except for the pollution sensitive taxa, and that the extrapolation of ecological models for sensitive taxa to another river basin with similar climatic and environmental conditions is more easy than for pollution insensitive ones. The main reason for this might be that these sensitive taxa have clear conditions in which they are present, whereas pollution insensitive taxa might have a diverse set of adaptive strategies, leading to different responses. Authors stressed that standardized data collection methods and data sharing are crucial for a convenient development of such international models, and that this is currently one of the major bottlenecks for applied environmental modeling research in general.

The paper *A rotation-invariant feature space according to environmental applications needs in a Data mining system using fish otoliths* is in the field of data mining systems that are feeded by images. Many environmental applications are related with the availability images as main source of information. The paper presents a Data Mining System for classifying fish otoliths, based on classifier methods. The input of the system are the query images of otoliths, which, as usual with images, present an arbitrary rotation. Feature extraction from images to find shape descriptors is an open issue due to this arbitrary rotation of the images, and accurate normalization is required to extract information from them, relevant for the data mining process. In this case, intensive feature extraction process in the preprocessing step is required to transform images into a set of discriminant features that can be in-

troduced into the classifier. A Main contribution is an innovative approach to preprocess the original images by producing a discriminatory feature space that describes the intrinsic characteristics of the otolith, independently of the rotation occurring in the image.

The paper *Efficient automated quality assessment: Dealing with faulty on-line water quality sensors* is in the field of data quality in water quality monitoring from data comming from on-line sensors. On-line monitoring systems and in situ sensors are increasingly used in the environmental sector. However data quality is instrinsically limited by technical reasons related with the performance of sensors in environmental contexts. Data quality assessment of water quality measurements is still an open issue because of the properties of water quality parameters themselves, and it is still manually treated. The paper presents a novel approach for a consistent and reliable water quality monitoring strategy for water systems. Automatic data quality assessment of water quality time series is proposed, based on exponential smoothing models and a dynamic prediction interval, which are combined to detect doubtful and unreliable data from raw complex data set. A posterior analysis is applied to remove noise through a kernel smoother and statistical features are then calculated to detect abnormal situations and potential sensor faults. The proposal is implemented as part of the RMS30 software and used within monitoring stations in several water systems (after proper calibration) like sewers, wastewater treatment plants (WWTP) and river bodies among others. The resulted validated data has many potential applications including the identification of dynamics, study of cause-effect relationships, capture sudden pollution events, resource management, modelling and integrated real time control of the catchment.

The paper *Development and selection of decision trees for water management: impact of data preprocessing, algorithms and settings* is in the field of ecological surface water quality by means of classification and regression trees. The work uses physical-chemical conditions and ecological conditions (macroinvertebrate community) to assess wether Flemish rivers approach the Water Frame Directive (policy making support), and providing a support to decide allocation of resources for river restoration. A methodological contribution of the paper is the analysis of the effect of preprocessing and model parameterization on classification and regression trees models. The paper illustrates its consequences on the applicability of the models for end-users. They concluded that data stratification, number of cross-validation folds and pruning cannot only impact the trees' reliabilities, but also in particular alter the applicability of environmental models. The authors recommend that environmental modellers should make use of an exhaustive list of model parameterizations to develop and compare environmental models in a transparent and objective manner, that combine different evaluation indices with expert knowledge (both system as user experts). Consequently, general guidelines derived from their research may help modellers to develop reliable, stable and reproducible models and to efficiently select statistical and ecological relevant models that are meeting the needs of users.

The paper *SAX-Quantile based multiresolution approach for finding heatwave events in summer temperature time series* introduces SAX-Quantile (SAX-Q) method to identify heatwaves in climate. SAX-Q is an innovative extension of the Symbolic Approximation algorithm for sequential pattern mining. SAX-Q is more robust than SAX, as it uses quartiles instead of averages, but provides the possibility to combine temporal information at several levels of granularity for a better understanding of dynamic patterns. A heuristic criteria is developed to identify heatwaves from the SAX-Q patterns identified at daily and weekle resolution intervals. The method is universal (invariant to the average temperature) so it can be adopted to analyse how heatwaves evolve at any place worldwide and any clima. It is applied to identify heatwave events from 50 years long 8 hours temperature time series. It shows that heatwaves are becoming more frequent and more intense along time. Most importantly, the intensity of the heatwaves is growing much more rapidly than the rate at which summers are increasing in average temperature. A very interesting characteristic of this work, is that if flows towards the concept of integral data mining system, covering the whole Knowledge Discovery Process, from the very early data collection and preparation, to the very last step of post-processing and knowledge production to support further decision making. In the paper, a new data base containing one heatwave per register with all relevant information (dates, temperatures....) is built for further analysis, but it is also visualized in a comprehensive way for a wide range of end users, from environmental engineers to policy makers. An application to London city (UK) is presented.

The paper *Use of symbolic dynamic time warping in hierarchical clustering of urban fabric evolutions extracted from spatiotemporal topograhic databases*

is a contribution for a better management of natural resources and territories (landscape). It provides a methodology for a better understanding of how urban areas change over time, also relevant in urban planning and management as well as to understand related antrophic processes. The paper tackles the spatio-temporal nature linked to urbanisation phenomenon. Spatio-temporal structure is a challenging characteristic of environmental systems and simultaneously dealing with both space and time is one of the hot topics from the data mining point of view. The work is based on a spatio-temporal clustering process that finds understandable temporal patterns of urban sprawl or its densification. The input of the process comes from topographic databases providing a description of elementary urban objects (mainly buildings) together with their infrastructures. There is one topographic database per year, describing the target area (a city...). At a first step, urban blocks are identified by using communication objects available in the topographic data base. Blocks are classified by using morphological features, according to a pre-existent set or patterns which are semantically interpreted in terms of urban configurations. The obtained labels are used as the basic vocabulary to build temporal sequences of urban block configurations along time, which are in turn clustered to find patterns of urban evolution. The principal originality of this approach is to use a distance measure based on DTW (Dynamic Time Warping) which is able to apprehend temporal behaviors (mainly time lags in dates corresponding to a change of state). The proposed distance takes into account the semantic proximity between the different kinds of urban blocks. In this work, semantic is introduced in the distances by using specific domain knowledge in form of an expert-based quantification of the similarity between the different types of urban blocks. An application on areas in the city of Strasbourg (France) in presented. Understandability of discovered patterns is also stressed in this paper.

The paper *Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sectors* identifies the major challenges facing decision-making for the urban water system sector and asserts that decision support systems are necessary in confronting these quandaries. The paper then reviews the literature to determine the extent to which machine learning methods used in data mining are applied to urban water systems' decision making. The application of these methods to decision support systems is found to be still limited and some hypothesis about reasons for this are identified.

These eight papers also show some sinergies among them, giving a wider view of the possibilities of the data mining in environmental sciences.

The data quality assessment on sensor data provided in [1] is valid for other water quality systems, like the ones presented in [2] or [4], but also valid in other environmental systems, like in air quality monitoring systems or weather sensors like the ones used in [8]. Data validation also constitutes a key module from decision support systems (DSS), like the ones referred in [7]. These data quality methods could be complementary within noise removal, data compression and event's identification approaches. Monitoring systems are used to raise alarms on abnormal behaviours, whereas [1] is using prediction intervals on upper and lower water limits to detect both sensor failures or water quality problems, [8] is using upper and lower quantiles to detect heatwave events, both works providing data-driven contributions to the classical idea of out-of-control charts in a context where observed phenomenons are non stationary, and follow complex patterns.

The preprocessing and parameter setting guidelines provided in [2] could be adopted by all works regarding classification and regression trees, like [9].

Data sharing policies proposed in [4] could also be used to build an international database, for example an international register of heatwaves [8].

Both [8] and [5] are managing the complex temporal structure of both temperature series and urban blocks evolution in cities by means of the use of sequential pattern mining methods providing a symbolic data representation for temporal patterns and using text mining techniques to find the temporal patterns. This approach is used in [5] on clustering spatiotemporal databases, while it is the basis for the time series analysis proposed in [8]. Still more, the dynamic time warping employed at [5] paper is one of the tools used in SAX literature to carry out comparison of different time series.

Even if both the Data Mining field and the Decision Support Systems field are mature research areas sufficiently developed to be extensively applied for a wide variety of Environmental problems, the survey presented in [7] concludes that there is still no wide use of data mining in decision support for environmental systems, and points out some reasons for that. As said before, one might be related to the lack of knowledge of the potential contributions of data mining methods to the environmental fields. Another one, might be the difficulty to understand row data mining

results from the environmental scientists. In this sense, works stressing the interpretability of the data mining model, or the inclusion of prior domain knowledge in any of its forms, may contribute to bridge this gap. This issue contains some works with a global decision support goal that address some of this issues in different ways. We expect that this special issue provides a wide view that permits to strength the sinergies between these areas and opens the possibility to tackle current environmental challenges in a successful advance of the current state of the art.

The elaboration of this issue counted with the unvaluable support of a Scientific Committee including 17 members from universities and enterprises from UK, France, Belgium, Denmark, Spain, Romania, Brasil and Australia. Nine of them belong to the data mining field, the other eight to the environmental sciences field. Five of them with active multidisciplinar research in both areas. Each paper has been reviewed at least by one specialist in data mining and one specialist in environmental sciences. I want to especially thank the Scientific Committee for the excellent and agile reviewing work done.

I would like to express my gratitude to all the authors who supported this special issue by sending their contributions and hope it provides valuable materials for a better understanding on how data mining can contribute to provide useful knowledge in environmental sciences. I also wish this issue becomes fruitful for researchers and constitutes inspiring materials to advance the state of the art in environmental data mining and to promote the use of data mining in environmental sciences.

References

[1] J. Alferes and P.A. Vanrolleghem, Efficient automated quality assessment: Dealing with faulty on-line water quality sensors, *Artificial Intelligence Communications* **29**(6) (2016), 701–709.

[2] G. Everaert, I. Pauwels, E. Bennetsen and P.L.M. Goethals, Development and selection of decision trees for water management: Impact of data preprocessing, algorithms and settings, *Artificial Intelligence Communications* **29**(6) (2016), 711–723.

[3] U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, Vol. 21, AAAI Press, Menlo Park, 1996.

[4] M.A.E. Forio, W. van Echelpoel, L. Domínguez-Granda, S. Tiku Mereta, A. Ambelu, T. Huong Hoang, P. Boets and P.L.M. Goethals, Analysing the effects of water quality on the occurrence of freshwater macroinvertebrate taxa among tropical river basins from different continents, *Artificial Intelligence Communications* **29**(6) (2016), 665–685.

[5] P. Gançarski, A. Puissant and F. Petitjean, Use of symbolic dynamic time warping in hierarchical clustering of urban fabric evolutions extracted from spatiotemporal topograhic databases, *Artificial Intelligence Communications* **29**(6) (2016), 733–746.

[6] K. Gibert, M. Sànchez-Marrè and J. Izquierdo, A survey on preprocessing techniques: Relevant issues in the context of environmental data mining, *Artificial Intelligence Communications* **29**(6) (2016), 627–663.

[7] A. Hadjimichaela, J. Comas and L. Corominas, Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector, *Artificial Intelligence Communications* **29**(6) (2016), 747–756.

[8] M. Herrera, A.A. Ferreira, D.A. Coley and R.R.B. de Aquino, Sax-quantile based multiresolution approach for finding heatwave events in summer temperature time series, *Artificial Intelligence Communications* **29**(6) (2016), 725–732.

[9] P. Martí-Puig and R. Reig-Bolano, A rotation-invariant feature space according to environmental applications needs in a data mining system using fish otoliths, *Artificial Intelligence Communications* **29**(6) (2016), 687–699.