International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

A picture on Environmental Data Mining Real Applications. What is done and how?

Karina Gibert[a,b], **Miquel Sànchez Marrè**[a,c]
[a]*Knowledge Engineering and Machine Learning group. Universitat Politècnica de Catalunya-Barcelona**Tech,** Spain*
[b]*Dep. Statistics and Operations Research. Universitat Politècnica de Catalunya-Barcelona**Tech,** Spain*
[c]*Dep. Software. Universitat Politècnica de Catalunya-Barcelona**Tech,** Spain*

Abstract: In this work a proposal for making systematic state of the art is presented and applied to the Environmental Data Mining field. The main characteristics of the Data Mining process have been identified. A form has been created to check which of those characteristics take place in a real application and how. A random sample of Science Citation Index papers regarding Data Mining and Environmental Applications has been selected. Papers were read by a set of experts and a form was filled in for every paper. The resulting information was mined itself using basic statistical analysis and some specific treatments for multi-response variables, to get a first picture of what is currently being done in the applications of Data Mining methods to environmental fields. Very interesting results have been obtained which depict very useful information. This information ranges from a general picture of what kind of methods are commonly used to which environmental areas seems to be more deeply using the data mining techniques. The paper presents and discuss these results, together with a proposal for building a continuous collaborative pannel in the web for enlarging the sample of papers and update the picture continuously. This will be easily possible because the analysis of the recorded data has been automatized in a statistical package set of macros for repetitive updating mined knowledge. The proposal is oriented to provide an Environmental Data Mining Observatoire, where getting updated information on what is being done in the area, identify drawbacks, orient future research in the methodological field to provide answer to the open environmental problems and finally, to give the environmental audience a wide corpus of previous experiences to be used as a reference for new applications.

***Keywords**:* Data Mining, state of the art, survey, form, observatoire

1 INTRODUCTION

Environmental Data Mining is a rather young research area were Data Mining (DM) tools are exploited to better analyse and understand environmental processes and systems. However, it is not well-known yet how environmentalists use Knowledge Discovery (KDD) techniques. For this purpose, a systematic survey of what is currently done in the area has been performed, in order to get a first accurate picture.

In a first step, the main characteristics of a DM process have been identified and a specific form has been designed to get relevant information from the literature. As a

first approach, the search has been limited to journals included in the Science Citation Index database, from Thompson Reuters and a set of experts read the papers and filled-in the forms, which were mined to give a first overview of the area.

The methodology used in this paper seems a sound way to perform state of the art and our idea is to provide a web-service to the general audience for both enlarging the database of analyzed papers and being updated in the area in an easy and collaborative process.

In the paper, section 2 provides information about the sample analyzed, section 3 provides the form used to synthetize the considered papers, section 4 provides the results of the analysis and section 5 some conclusions.

2 THE SAMPLE

The papers found containing at least one keyword related to Data Mining and one related to Environemental domains were 3015:

> *Search for papers with a list of keywords such that: Topic=("data mining" or datamining or "Knowledge Discovery" or KDD) AND Topic=(Environment* or ecolog* or agricult* or water or soil or forest or pollu* or climate or air or land)*

From those, only 2163 papers published in English were retained and 1439 appeared in conference proceedings or journals too far from our target, like 7TH INTERNATIONAL SYMPOSIUM ON TEST AND MEASUREMENT, or JOURNAL OF SPACECRAFT AND ROCKETS, and were not included in this study. From the remaining 724 papers a random sample of 46 papers was selected and carefully analyzed by a team of 7 experts from both Data Mining (DM) and Environmental domains. Figure 1 illustrates the sampling process.

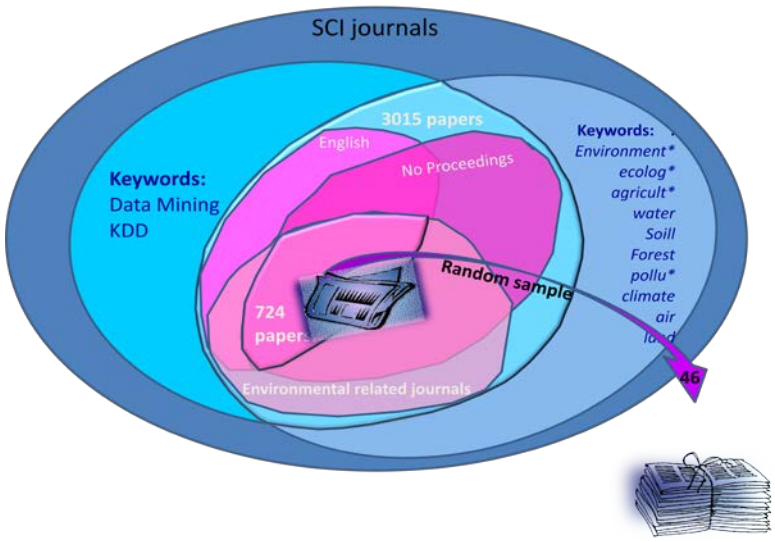

Fig. 1. The sample

3 THE QUESTIONAIRE

Traditionaly, the state of the art has been done by reading the papers and extracting the relevant characteristics of all of them, till some general conclusions can be elaborated on the basis of the expertise of the reader. In this paper, we propose a more structured process based on a previous analysis to identify which

are the relevant characteristics of the target field. Those characteristics can be presented to the reader or the researcher in form of template to be filled-in, and to provide structure to the information extracted from the paper. This will permit to place the whole consulted works in a standard frame and will make easier the comparisons and getting a global perspective.

That's why for every paper a brief form was filled-in, specifically designed to get relevant information about the data mining methods used, the kind of data treated, the kind of methodology used or other relevant aspects of the data mining process pursued in every paper (Fig. 2). The contents of the form considers all the aspects of a data mining process:

- Structure of the data
- Preprocessing
- Data Mining
- Post-processing
- Validation
- Software used

ID data Author: _____ Journal: _____ Year: ____

Environmental target field Water, Ecology, Land, Agriculture, Forest, Air, Climate, Other (circle or specify) _____

Main goal of the paper Descriptive ☐ Predictive ☐

Types of data

Numerical ☐ 2D ☐ Temporal ☐ Various granularity ☐

Qualitative ☐ 3D ☐ Various periodicity ☐

Database size #Objects: _____ #Attributes: _____

Preprocessing Addressed? (Yes/No) #used techniques: _____

Missing data ☐ Variable selection ☐ Feature Weighting ☐

Variables Transformation ☐ Dimensionality reduction ☐

Other (specify) _____

Use of previous knowledge No ☐

Rules ☐ Constraints ☐ Ontology ☐ Previous distribution ☐

Other (specify) _____

DM techniques #Used techniques _____

Hybridation ☐ Sequential combination ☐ Separate use for comparison ☐

Clustering ☐ Ass-RuleInd ☐ Mvar ☐ BN ☐ CBR ☐ Class-RuleInd ☐

DecTrees ☐ RegrTrees ☐ SVM ☐ ANN ☐ DiscrAn ☐ StatRegr ☐

ANOVA ☐ TimeSerAn ☐ Other (specify) _____

Association-Rule Induction; Multivariate Analysis; Bayesian Networks; Case Base Reasoning; Classification-Rule Induction; Decision Trees; Regression Trees; Support Vector Machines; Artificial Neural Networks; Discriminant Analysis; Statistical Regression; Time Series Analysis

Post-processing Addressed? (Yes/No) Visualization ☐ Other (specify) _____

Validation NA ☐ No ☐

Randomization ☐ Cross validation ☐ Benchmarking ☐ Visualization ☐

Other (specify) _____

Software tools #Used tools _____

Weka ☐ R ☐ CART ☐ Clementine ☐ LIBSVM ☐

Self-developed (specify) _____ Other (specify) _____

Fig 2. Form used to analyze the contents of the reviewed papers

3 THE ANALYSIS

The contents of the form were inputted in a statistical package and some descriptive statistics performed. The results are briefly described below.

Papers explicitly talking about Data Mining or KDD starts from 1997 on, according with the foundation of the discipline in 1996; the greater production is found arround 2005. A qualitative increasing of papers in the area is found from 2004 on. This confirms that area is still active (fig 3 left)

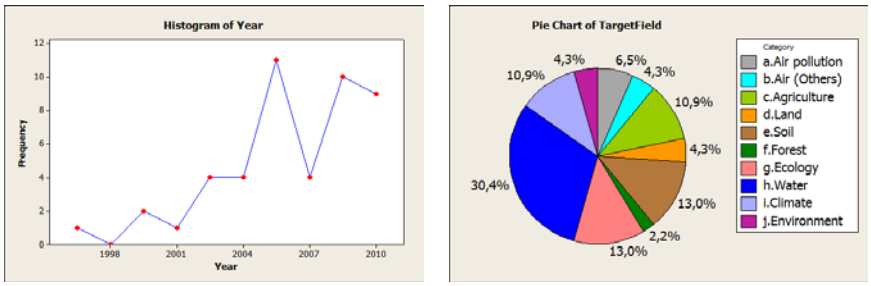

Figure 3: (left) Year of publication; (right) Target field

Most of the applications regards water (Fig. 3 right). About half of the papers pursue descriptive goals (47,83%), and the other half predictive goals (52,17%). No preponderance of predictive models, in contrast of classical statistical analysis.

The dataset size involves both the number of variables and the number of instances. The number of instances can be important in some of the works (50% of papers work with more than 950 objects and 25% with more than 4859 objects. But less than a 25% include a really huge number of instances). But less than a 25% include a considerable number of instances and none of them are in the orders or terabytes or higher. The number of variables considered varies from 5 to 147 variables, which is rather far from massive datasets.

Most of the applications involve numerical data, only 4,3% didn't. A 19,56% of the papers treat simultaneously numerical and qualitative data and a 89,13% involve heterogeneous data (including numerical and qualitative or streams or graphical data). They regard to agriculture, land, forest, water or climate.

An important characteristic is that space (52,2%) or time (34,8%) use to be present in environmental applications, in front of a 13.3% and 38.3% respectively reported for general purpose data mining applications in [Kdnuggets 2007]. In fact, more than a 80% of the applications reviewed include temporal or spatial information (figure 4). But only 21,74% papers involve both simultaneously. We think that this is more related to a lack of appropriate tools to catch the whole complexity of such studies. On the other hand, surprisingly, among the applications involving time, only 16,6% use time series methodologies.

The coexistence of different scales of spatial or temporal data is also a characteristic of the area. A 43,5% of papers involve at least one of those. This can have implications on sparsity of data matrices and false missing data, which require special treatments that not always are taken into account.

Surprisingly 10,8% of papers involve genetic data, which was not expected in ES. All of them are related with agriculture or soil except one which is targeted to air pollution. This indicates that, even with a marginal presence, in some applications genetic data makes also sense in ES and as this kind of data is considered, special data mining techniques will become required for treating them.

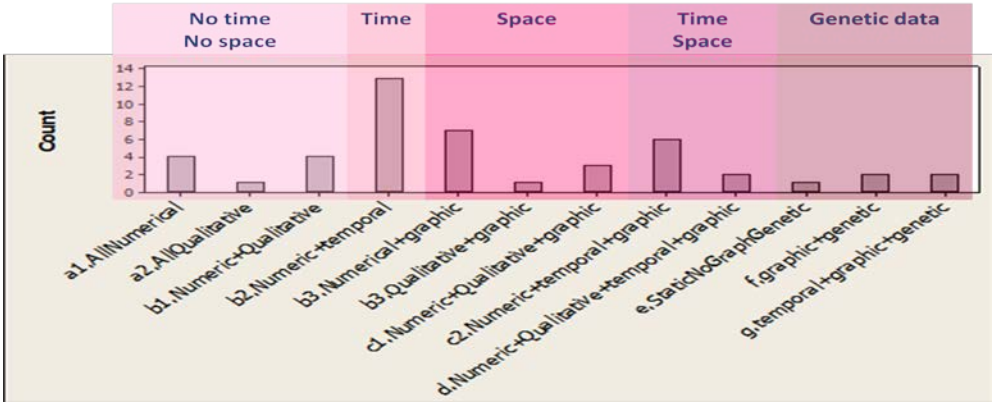

Figure 4: Types of datasets in environmental data mining applications.

Less than a 25% of works take into account existent prior expert knowledge; this may indicate that these possibilities are not well known in environmental sciences (ES) yet. However, when prior expert knowledge is introduced into the analysis, formal knowledge models seem to be preferred, like ontologies (45,5%) or rule-bases (26,6%), and only exceptionally (17,2%) the knowledge is directly acquired from the experts.

Half of the papers use a single data mining method, but the rest use between 2 and 9 in different ways. Mainly to perform inter-method comparisons (56,52%), less often to combine sequences of methods in a complex data mining process (34,7%) in such a way that the output of a method is the input of a subsequent one, and rarely hybridation of methods in a more powerful method is introduced (14,04%).

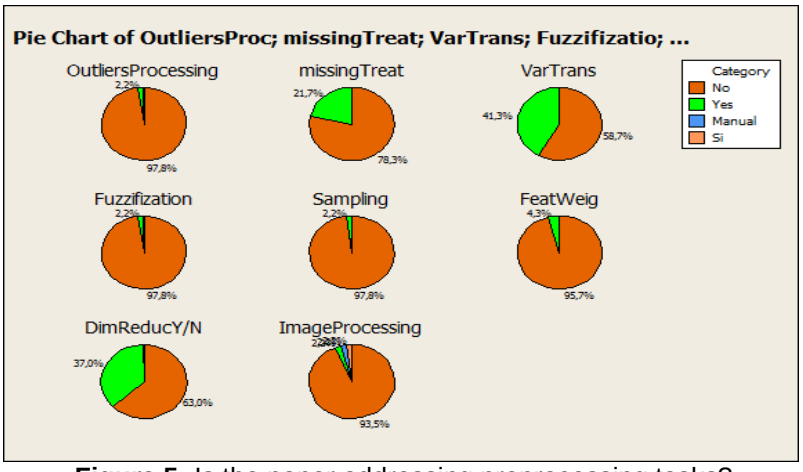

Figure 5: Is the paper addressing preprocessing tasks?

The 76,09% of the papers address preprocessing skills, using between 1 and 5 methods for that, from those reported in Fig. 5..

The most used DM technique is the decision tree (26,09%, usedful in predictive problems) followed by Clustering (23,9%, useful in descriptive problems) (Fig 6).

A 26,09% performs data mining using only statistical methods like multivariate analysis, general linear models (regression, ANOVA....) or time series analysis; while a 36,96% use only AI methods like classifiers, association rules, Bayessian networks, case-based reasoning, regression trees, support vector machines or artificial neural networks (this mainly for predictive purposes, altough a couple of papers introduce them for non-supervised analysis).

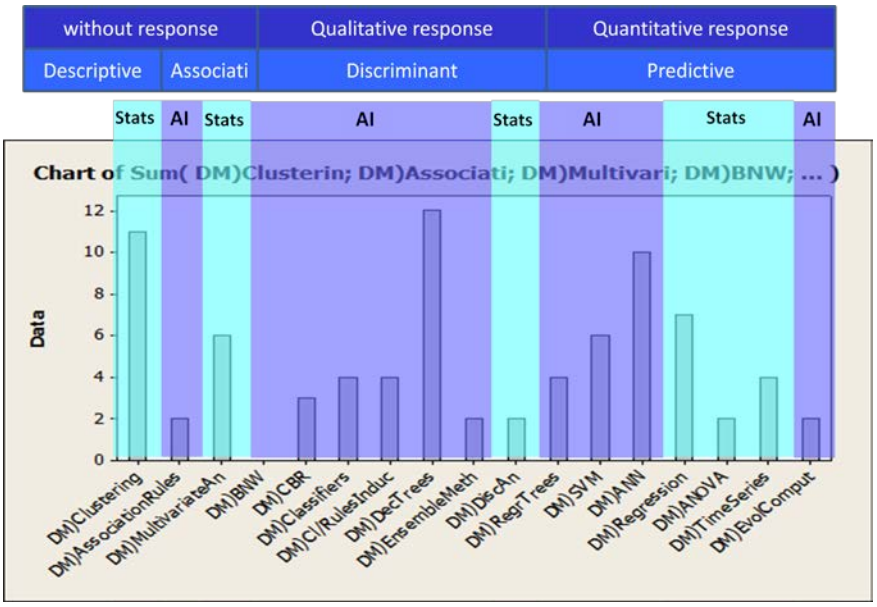

Fig 6. Data Mining techniques used

Qualitative DM methods seem to be preferred in real applications, probably because they produce directly understandable results, which enormously improve the links between the results of the DM process and the decision-making. In fact, understandability and usability are key points in DM. That is why, postprocessing is critical particularly for numerical methods with non intuitive results.

Unfortunately, postprocessing is method-dependent and seems to be much underdeveloped. Only a 52,17% of the papers addresses it. In fact, from those, a 79,2% only use visualization techniques to represent results. Marginally, other postprocessing techniques are used, like comparison, ranking of multiple solutions, manual interpretation of results, expert-guided refinement of solutions, automatic refinement of solutions like pruning through quality indicators, tranformation of variables for reanalysis, or induction of an interpretative model for results (like production rules analysis for decision trees or induction of automatic interpretation of profiles for clustering) . A high degree of heterogeinity is found on postprocessin and global guidelines seem to be required.

Validation is still method-dependent and 65% of papers address it. The remaining 35% refer descriptive analyses or global environmental theoretical approaches. Among the most used validation strategies: simple or cross-validation (36,95%). Only marginal use of other validation strategies appears: quality index evaluation (8,7%), multimethod comparison (6,5%), expert validation (6,5%), comparison with a reference variable, benchmarking or sensitivity/specificity analysis.

A 47,8% of papers involve the use of a single software, but there are some papers that involve up to 5 different software tools, mainly oriented to perform comparisons. Around half of the applications apply softwares working with standard rectangular data matrices with crisp data. But a number of applications involve images, or geospatial data or data streams.

Surprisingly (Fig. 7), the most famous commercial Data Mining packages have not been often used in those researches, even if they include most of the data mining methods used. Authors guess that standard non-research applications rarely provide methodological contributions and are rarely published in research papers, which makes difficult to get knowledge about them.

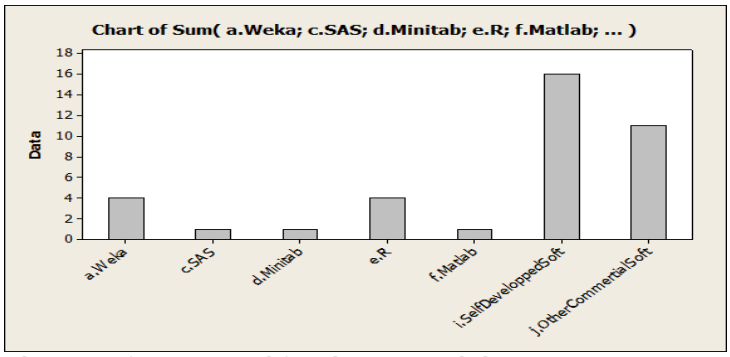

Fig. 7: Software used for the Data Mining

5 CONCLUDING REMARKS

KDD, often simply known as Data Mining, defines a new paradigm where, apart from the data mining step itself (devoted to data exploitation), data preparation or results post-processing are included in the methodology itself, as well as the interest to take advantage of the existent prior expert knowledge to improve usability of results.

The DMTES workshop series started in 2006 with the aim of becoming a meeting point between environmentalists and data miners and a double purpose: On the one hand, to disseminate the possibilities of the DM techniques to contribute to better understand, predict, control and manage environment in general; on the other hand, to provide to data miners a clear insight to real environmental problems and open research opportunities oriented to environmental fields. After some editions, in [Gibert 2011] we described the main outcomes extracted from the workshop and session series.

The survey presented in this paper was performed with a primarily aim of checking whereas the DMTES series of workshops and sessions can be followed as a reliable observatoire of how environmental data mining evolves in general.

It is highly significant to verify that many of the trends observed from the DMTES series contributed papers have been found again in a non-oriented survey using a completely different information sources, the SCI journals database. This a strong point to consider the DMTES-Workshop series, held in the context of iEMSs conference, as a reference to be updated on what is happening in the area.

One of the most interesting findings is that many environmental DM applications do not necessarily imply massive data sets analysis. In fact, big data with huge number of objects require highly efficient algorithms, while huge number of variables is more related with a need of powerful models that can properly deal with complex interactions between those variables. In most applications analyzed, the data mining approach is more used to cope the whole complexity of the problem than to deal with massive datasets, what is aligned with the new possibilities of DM to model more complex phenomena or systems.

The pretreatment procedures used include all those announced in [Gibert et a 2011l]: Outliers detection and treatment, Missing data treatment, Variables transformation, Dimensionality reduction (either from rows or columns of the data matrix, in different ways), including feature weighting techniques which do not directly reduce the dimensionality. However, only missing data, variables transformation and dimensionality reduction show a frequent use in the analyzed applications. We expected a higher use of outliers treatment. Feature weighting shows a marginal use, maybe related with a lack of habits to use the prior knowledge to decide feature weights, or with the lack of popular software tools that can easily incorporate those weights into the analysis.

There are few issues found which weren't mentioned in [Gibert et al 2011] and that were used in a 8,6% of the papers:

- Image preprocessing for extracting relevant features from images to build the data matrix for the analysis. Of course related with geospatial and signal processing applications
- Fuzzyfication of originally crisp data to use fuzzy data mining methods, in an agricultural application
- Sampling on the rows of the original data matrix to reduce the dataset

More complex data mining processes seems to increase over time, according to the idea that more and more complex environmental processes are trying to be modelled . However, even if it starts to be frequent to use several DM methods in the analysis, it is still marginal to use new hybrid methods that could intrinsically capture more complex domain structures. Research in this field is still young.

A considerable number of papers (28.26%) show the multidisciplinariety typical in the area, using both statistics and AI methods in the data mining process . However, as shown in fig.6 only few techniques are really popular in environmental sciences (ES) and a number of available methods are marginally used in real applications. Also, in a number of cases, predictive techniques have been applied to applications with descriptive goals, which can indicate some confusion in the area on the technical point of view. In that sense, the methods' conceptual map presented later in this paper can help to improve choice of more appropriate technique in front of a real problem [Gibert et al 2010].

The presented survey provides a preliminary overview of what is being published on environmental data mining. This has now to be related with how Data Mining is used out of research fields. A second wave of papers is currently being analyzed under the same scheme. With bigger set of data second order information could be extracted from data going further than basic descriptive statistics. Also, a service is being prepared to enable the use of the questionnaire to the general audience. In the near future environmentalists or data miners reading papers, will be allowed to fill-in the form via a dedicated web site, in such a way that the survey could be periodically updated in a collaborative Observatoire of Environmental Data Mining.

ACKNOWLEDGMENTS

Specially to Joaquín Izquierdo, Geoff Holmes, Serena Chen, Ignasi Rodríguez-Roda, t oread part of the papers and fill-in the questionaires

REFERENCES

Gibert, K., Sànchez-Marrè, 2011. Outcomes from the iEMSs Data Mining in the Environmental Sciences Workshop Series. Env. Modelling and Software, 26:983-985

Gibert, K., Sànchez-Marrè M., Codina V. 2010 Choosing the right data mining technique: classification of methods and intelligent recommenders. In Proc. of the iEMSs 5th Biennial Meeting vol. I, 1933-1940 Swayne, D. et alt., Ottawa University, CA.

KdNuggets 2007 http://www.kdnuggets.com/polls/2007/data_types_analyzed.htm