International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-7, May 2021

Survey of Process of Data Discovery and Environmental Decision Support Systems

Altaf Alaoui, Boris Olengoba Ibara, Badia Ettaki, Jamal Zerouaoui

Abstract: ***The process of data discovery is an approach to extracting knowledge, valid, and usable information from large amounts of data, using automatic or semi-automatic methods. This article is an inventory of the different information extraction processes encountered in the literature for different fields of application and for the development of environmental informatics. Following an analysis between the different models, we can summarize the existing models with a proposal for a process that exploits the strengths of the different processes.***

Keywords: Knowledge Discovery in Databases, KDDM, Environmental Decision Support Systems, GIS, KBS, EDSS.

I. INTRODUCTION

The literature often uses the term KDD "Knowledge Discovery in Databases" to refer to the process of data discovery. The KDD's objective is to understand the need, expressed by the expert, to provide understandable, useful results and avoid the indiscriminate application of analytical methods 1 2 3. In addition, it also contributes to good management and work planning, to ensure and guarantee good communication, especially within relatively large work teams. The development of KDDs is adapted to the scope of application and the nature of the data. In the context of environmental problems, the terms "Environmental Decision Support Systems" (EDSS) 4 or "Multiple Objective Decision Support Systems" (MODSS) 5 are generally used, to refer to the process of data discovery. These are intelligent information systems that improve the time it takes to make decisions, as well as the consistency and quality of decisions, expressed in quantities characteristic of the scope.

This survey is based on a comparative study of the different processes of data discovery and their areas of application, a study of environmental decision support systems, and the proposal of a process that unites the strengths of the processes studied, and a conclusion.

II. COMPARATIVE STUDY OF THE DIFFERENT DATA DISCOVERY PROCESSES

The KDDM "Knowledge Discovery and Data Mining" processes of Fayyad et al. 2 and the CRISP-DM "Cross Industry Standard Process for Data Mining" 6, are the most encountered in the literature, and are considered the de-facto standard or main matrix in the KDD field. As a result, several other processes are derivatives of both models.

The first model of KDDM reported includes nine steps 2:

1) Understanding the scope: During the first step, the objectives are defined from the customer's perspective and used to develop and understand the field of application;
2) Creating a target dataset: In the second step, the target dataset will be created;
3) Data cleaning and pre-processing: The main objective is to clean up and pre-process the 'target' data to obtain complete and consistent data;
4) Data transformation: This is about transforming data from one form to another so that data mining algorithms can be easily implemented. Hence, different methods of data reduction and transformation are implemented on the target data;
5) Choosing the appropriate task of data mining: The objective of the fifth step is to choose the appropriate data mining task based on specific objectives defined in the first stage. Examples of data mining methods or tasks are classification, grouping, regression, synthesis, etc.;
6) Choosing the appropriate algorithm: One or more appropriate data mining algorithms are selected to search for different models from the data. The selection of the right algorithms is based on the corresponding criteria for data mining;
7) Algorithm use: Consists of implementing selected algorithms;
8) Interpretation of the results: This involves interpreting, evaluating, and visualizing the models extracted;
9) Use of discovered knowledge: This is the stage at which the discovered knowledge is used for different purposes.

The KDDM process model has been applied in several areas such as, Medicine, Engineering, Industrial production and regulation, E-commerce, Software development, etc.

Manuscript received on May 06, 2021.
Revised Manuscript received on May 19, 2021.
Manuscript published on May 30, 2021.

* Correspondence Author

Alaoui Altaf*, Laboratory of Materials Physics and Subatomics. Faculty of Sciences- Ibn Tofail University, Kenitra, Morocco. Email: altaf.alaoui@gmail.com

Boris Olengoba Ibara, Laboratory of Ecology and Environment, Faculty of Sciences Ben M'sik, University Hassan II, Casablanca, Morocco. Email: bobholens@gmail.com

Badia Ettaki, Laboratory of Research in Computer Science, Data Sciences and Knowledge Engineering , School of Information Sciences Rabat, Morocco, Email: ettakibadia@yahoo.fr

Jamal Zerouaoui, Laboratory of Materials Physics and Subatomics. Faculty of Sciences- Ibn Tofail University, Kenitra, Morocco. Email: j_zerouaoui@yahoo.fr

Several approaches which are directly derived from the basic KDDM process have been proposed, such as:

- Adriaans and Zantinge process model 7, which consists of six steps that correspond successively to data selection, cleaning, Data enrichment, Coding, Data Mining, and reporting;
- Berryet Linoff's process model 8 including four steps: problem identification, problem analysis, action taken and outcome evaluation;
- SEMMA "Sample, Explore, Modify, Model, Assess" model developed by SAS Institute Inc. 9, it includes five steps that correspond respectively to: sampling, exploration, modification, modeling and evaluation.

The process models mentioned above have been used in the industry. However, the literature does not specifically mention the type of industry.

In addition, there are other types of process models, which have been used specifically in the marketing and sales sector. These include:

- Cabena et al.10 process model, which is run in five stages: goal setting, data preparation, data mining, domain knowledge exploitation, and knowledge assimilation. This model is used more in industrial projects 11 12;
- Anand-Buchner 13 process model consisting of eight steps: human resource identification, problem specification, data mining, domain knowledge exploitation, method identification, data processing, model discovery and post-processing of knowledge.

And some process models have been developed and applied both in research fields oriented towards industrial engineering and medicine.

These are the process models of:

- CRISP-DM, designed for business project management, which consists of six steps including, understanding objectives, understanding data, preparing data, modeling, evaluation, and deployment. It is a model that has been used for data extraction for industrial engineering and medical research projects 14 15 16 17 18 19. The CRISP-DM was also used in industrial projects (marketing and sales) 20 21 22 23 24 25 26 27.
- Cios et al. process model was proposed for the first time in 2000 28 29, an adaptation of the CRISP-DM model aimed this time at meeting the needs of the university research community (Medicine, Software) 30 31 32 33 34 35 36 37 38 39.
- Han-Kamber process model 40, which is summarized in seven steps. Successively, it includes mastering the problem, creating a target dataset, cleaning and pre-processing data, reducing and transforming data, selecting and organizing the necessary personnel and tools resources, choosing the data mining algorithm or algorithms, choosing the right tool for data management, evaluating models and presenting knowledge, and use of the discovered knowledge;
- Edelstein's process model 41, condensed into five steps, namely, problem identification, data preparation, model construction, model use and model evaluation;
- Klosgen-Zytkow process model 42, eight steps compose it: defining and analyzing business problems, understanding and preparing data, setting up knowledge research, finding knowledge, refining knowledge, applying knowledge to meet business needs, deploying and evaluating practical solutions;
- Haglinet al. 43 model, runs in seven steps: objective identification, targeted data creation, data pre-processing, data transformation, Data Mining, evaluation and use of extracted knowledge.

By examining the different models of data mining processes, they clearly show that they have strong similarities to each other. Several process models have kept the main features of the KDDM, and the fact that they have in common most of the main steps corroborates this finding. Moreover, most have been applied for the needs and problems inherent in business and rarely for the purposes of scientific research.

In the latter case, only the work associated with the management of scientific projects used it. However, given the wide range of data available from the various sparse databases dedicated to monitoring nature and the environment, the establishment of a tool such as a KDDM adapted to environmental data to extract knowledge for optimal use, would give a boost in research oriented towards management and conservation of the environment. To this end, a new research axis is emerging: environmental computing 44 45 46 47 48.

III. ENVIRONMENTAL DECISION SUPPORT SYSTEMS

The development of environmental computing is made possible with the profusion and availability of data that can produce relevant and useful information for understanding and managing the environment. Discipline combines research areas such as artificial intelligence, geographic information systems (GIS), mathematical models, modeling and simulation, etc. This new discipline develops appropriate methods for modeling, design, simulation, forecasting, planning and decision support systems for environmental management and protection.

The declination of artificial intelligence research towards environmental issues is illustrated by the development of knowledge-based systems "Knowledge Based System" (KBS) 50. These KBS systems, applied for this purpose, are often referred to as Decision Support Systems (DSS) 51 or Environmental Decision Support Systems (EDSS) 4.

The main characteristics of EDSS is the ability to extract and use knowledge from specialized and multi-source data associated with natural and environmental sciences.

Thanks to its features, an EDSS can address the following specific concerns:

1) Acquiring, representing and structuring knowledge (information) in the field of environmental study.
2) Possibility of producing portable sub-database models that can be used in environmental projects optionally dealing with different issues, which saves considerable time in the pursuit of current or future environmental projects.

3) Ability to integrate and process spatial data (the GIS component).
4) The ability to use multiple data sources to effectively diagnose, plan and manage the environmental components based on information quality optimization.

In essence, as a KDD, an EDSS can be described as a multi-layered system. Furthermore, the first layer of an EDSS is usually a knowledge acquisition and learning module from a spatial (GIS) and temporal (chronological and periodic) database. The next layer consists of several models of Artificial Intelligence, Statistics and Data Mining. The other levels are the subject of the reasoning and integration modules. These modules use several types of models and knowledge to provide prediction, description, planning or supervision information from the EDSS. Finally, the higher level illustrates the planning or supervision interaction from the EDSS.

However, according to the objective of the EDSS, two categories emerge 4. It distinguishes problem-specific EDSS and problem and location-specific EDSS. The former is adapted to relatively narrow environmental problems (or areas), but they apply to a wide range of locations (or situations), but the latter are adapted to both a specific environmental problem and a specific location.

IV. HYBRID PROCESS OF DATA DISCOVERY

A key feature of the KDD approach is the importance it gives to pre-processing phases (data selection, attribution of missing values, suppression of noise or outliers, mapping characteristic values on appropriate domains, etc.) to improve the quality of data contained in the dataset for the process to produce reliable knowledge. They are often specific to an area or a problem.

To this end, a comprehensive process can combine the benefits of existing KDD and EDSS, to strengthen the steps of information extraction and knowledge, from raw data, for decision-making assistance. The following steps summarize the KDD and EDSS unite:

1) Understanding the scope and its needs and the associated constraints.
2) Researching the different solutions adapted to the problem and the nature of the objects to be studied or of interest. This second step consists of a classification within the typologies and in-depth analysis of the different methods and algorithms that could constitute an adequate solution to the work context.
3) Choosing the appropriate solution: This step is the choice of the appropriate task, the latter depends on the compromise between the nature of the problem, the objectives assigned, and the features offered by different solutions that respond in part or in full to the needs of the study.
4) Data modeling and design: This step is dedicated to structuring data within the database.
5) Selecting and creating a target data set from the main database de-complexed in the previous step.
6) Acquisition of complementary data: In recent years several databases have been accessible (Open data), they make available to the general public, important data in research, including spatial data, precipitation data, temperature data, satellite images etc. This data is necessary to sharpen the accuracy and quality of the information to be produced.
7) Data pre-processing: The main objective of this step is to clean up and pre-process target data to obtain complete and consistent data.
8) Proposal, modeling exploration algorithms to meet study needs.
9) Implementation of algorithms: This is the stage of the process in which the proposed algorithms are exploited for the various predefined objectives;
10) Interpretation of results: This is the step in the process that focuses on interpreting and evaluating models of the knowledge exploited. This step may involve a visualization of the extracted patterns.

V. CONCLUSION

The aim of this paper is to offer an overview of the process of data discovery used for the different fields of application and for the development of environmental computing.

The survey also proposes to enhance existing models by unite the advantages of each process and especially the addition of EDSS processes which offers a layer dedicated to the exploitation of external data to strengthen the processing. This proposal is more suited to problems where the spatial aspect carries relevant information.

REFERENCES

1. Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and Ft. Uthurusamy, 1996. Advances in Knowledge Discovery and Data Mining, (AKDDM), AAAI/MIT Press.
2. Fayyad, U.M., Haussler, D. and Stolorz, Z. 1996. KDD for Science Data Analysis; Issues and Examples. Proc. 2nd Int. Conj. on Knowledge Discovery and Data Mining (KDD-96), Menlo Park, CA: AAAI Press.
3. Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. 1996. From Data Mining to Knowledge Discovery: An Overview, in AI(DDM, AAAI/MIT Press, pp. 1-30
4. D. A. Swayne, R. Denzer, L. Lilburne, M. Purvis, N. W. T. Quinn, and A. Storey, "?," in Environmental Software Systems, vol. 39, R. Denzer, D. A. Swayne, M. Purvis, and G. Schimak, Eds. Boston, MA: Springer US, 2000, pp. 259–268.
5. U. Baizyldayeva, O. K. Vlasov, A. A. Kuandykov, and T. B. Akhmetov, "Multi-Criteria Decision Support Systems. Comparative Analysis," 2013.
6. Shearer, C., The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22, 2000.
7. Adriaans. P and Zantinge.D, Data mining, Addison-Wesley, 1999.
8. Berry, M. J., & Gordon, L., Data mining techniques: For marketing, sales, and customer support. New York, NY: Wiley, 1997.
9. SAS Enterprise Miner – SEMMA. SAS Institute. Accessed from http://www.sas.com/technologies/analytics/datamining/miner/semma.html, on May 2008
10. Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A., Discovering data mining: From concept to implementation. Upper Saddle River, NJ: Prentice Hall, 1998.
11. Hirji, K. K., Exploring data mining implementation. Communications of the ACM, 44(7), 87–93. doi:10.1145/379300.379323, 2001.
12. Anand, S. S., Bell, D. A., & Hughes, J. G., The role of domain knowledge in data mining. In Proceedings of the 4th International Conference on Information and Knowledge Management (pp. 37-43), 1995.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

13. Anand, S.S., Büchner, A.G., Decision Support through Data Mining, FT Pitman Publishers, 1998.
14. Buchheit, RB, Garrett, JH, Jr, Lee, SR and Brahme, R, A knowledge discovery framework for civil infrastructure: a case study of the intelligent workplace. Engineering with Computers 16(3–4), 264–274, 2000.
15. Jensen, S., Mining medical data for predictive and sequential patterns: PKDD 2001. In Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2001 Discovery Challenge on Thrombosis Data, 2001.
16. Butler, S., An investigation into the relative abilities of three alternative data mining methods to derive information of business value from retail store-based transaction data. BSc thesis, School of Computing and Mathematics, Deakin University, Australia, 2002.
17. Blockeel, H. and Moyle, S., Collaborative data mining needs centralized model evaluation. In Proceedings of the ICML-2002 Workshop on Data Mining Lessons Learned, pp.21–28, 2002.
18. Silva, E.M., Do Prado, H.A. and Ferneda, E., Text mining: crossing the chasm between the academy and the industry. Management Information Systems 6, 351–361, 2002.
19. Jensen, S., Mining medical data for predictive and sequential patterns: PKDD 2001. In Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD2001 Discovery Challenge on Thrombosis Data, 2001.
20. Butler, S., An investigation into the relative abilities of three alternative data mining methods to derive information of business value from retail store-based transaction data. BSc thesis, School of Computing and Mathematics, Deakin University, Australia, 2002.
21. Blockeel, H. and Moyle, S., Collaborative data mining needs centralized model evaluation. In Proceedings of the ICML-2002 Workshop on Data Mining Lessons Learned, pp.21–28, 2002.
22. Silva, EM, Do Prado, HA and Ferneda, E., Text mining: crossing the chasm between the academy and the industry. Management Information Systems 6, 351–361, 2002.
23. Hipp, J and Lindner, G., Analyzing warranty claims of automobiles. An application description following the CRISP-DM data mining process. In Proceedings of 5th International Computer Science Conference, Hong Kong, China, pp.31–40, 1999.
24. Gersten, W., Wirth, R. and Arndt D., Predictive modeling in automotive direct marketing: tools, experiences and open issues. In Proceeding of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 398–406, 2000.
25. Moyle, S., Bohanec, M. and Ostrowski, E., Large and tall buildings: a case study in the application of decision support and data mining. In Proceedings of the ECML/PKDD'02 workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning, pp.88–99, 2002.
26. Li, S-T and Shue, L-Y, Data mining to aid policy making in air pollution management. Expert Systems with Applications 27(3), 331–340, 2004.
27. De Abajo, N, Lobato, V, Diez, AB and Cuesta, SR., ANN quality diagnostic models for packaging manufacturing: an industrial Data Mining case study. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 799–804, 2004.
28. Cios, K, Teresinska, A, Konieczna, S, Potocka, J and Sharma, S., Diagnosing myocardial perfusion from PECT bull's-eye maps—a knowledge discovery approach. IEEE Engineering in Medicine and Biology Magazine, Special issue on Medical Data Mining and Knowledge Discovery 19(4), 17–25, 2000.
29. Cios, K and Kurgan, L, Trends in data mining and knowledge discovery. In Pal, N and Jain, L (eds) Advanced Techniques in Knowledge Discovery and Data Mining. Springer, pp.1–26, 2005.
30. Sacha, J, Cios, K and Goodenday, L., Issues in automating cardiac SPECT diagnosis. IEEE Engineering in Medicine and Biology Magazine, Special issue on Medical Data Mining and Knowledge Discovery 19(4), 78–88, 2000.
31. Kurgan, L, Cios, K, Tadeusiewicz, R, Ogiela, M and Goodenday, L., Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine 23(2), 149–169, 2001.
32. Goh, KG, Hsu, W, Lee, ML and Wang, H., ADRIS: an automatic diabetic retinal image screening system. In Cios, K (ed.) Medical Data Mining and Knowledge Discovery, pp. 181–207, 2001.
33. Shalvi, D and DeClaris, N., A data clustering and visualization methodology for epidemiological pathology discoveries. In Cios, K (ed.) Medical Data Mining and Knowledge Discovery, pp. 129–151, 2001.
34. Cios, K (ed.) 2001, Medical Data Mining and Knowledge Discovery. Springer-Verlag.
35. Maruster, L, Weijters, T, De Vries, G, Van den Bosch, A and Daelemans, W, 2002, Logistic-based patient grouping for multi-disciplinary treatment. Artificial Intelligence in Medicine 26(1–2), 87–107.
36. Ganzert, S. Guttmann, J, Kersting, K, Kuhlen, R, Putensen, C, Sydow, M and Kramer, S., Analysis of respiratory pressure–volume curves in intensive care medicine using inductive machine learning. Artificial Intelligence in Medicine 26(1–2), 69–86, 2002.
37. Perner, P., Perner, H. and Muller, B., Mining knowledge for HEp-2 cell image classification. Artificial Intelligence in Medicine 26(1–2), 161–173, 2002.
38. Hofer, J. and Brezany P., Distributed Decision Tree Induction within the Grid Data Mining Framework GridMiner-Core. GridMiner TR2004–04, Institute for Software Science, University of Vienna, 2004.
39. Kurgan, L, Cios, K, Sontag, M and Accurso, F., Mining the cystic fibrosis data. In Zurada, J and Kantardzic, M (eds) Next Generation of Data-Mining Applications. IEEE Press and Wiley, pp. 415–444, 2005.
40. Han, J. and Kamber, M., Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.
41. Edelstein, H., Data mining: let's get practical. DB2 Magazine 3(2), summer, 1998.
42. Klosgen, W and Zytkow, J, 2002, The knowledge discovery process. In Klosgen, W and Zytkow, J (eds) Handbook of Data Mining and Knowledge Discovery. Oxford University Press, pp.10–21.
43. Haglin, D, Roiger, R, Hakkila, J and Giblin, T, 2005, A tool for public analysis of scientific data. Data Science Journal 4(30), 39–53.
44. Haagsma I.G. and Johanns R.D., "Decision support systems: An integrated approach," in Environmental Systems, edited by P. Zannetti, vol. II, pp. 205–212, 1994.
45. Gabaldo´n C., Ferrer J., Seco A., and Marzal P., "A soft- ware for the integrated design of wastewater treatment plants," Environmental Modelling and Software, vol. 13, no. 1, pp. 31– 44, 1998.
46. Guariso G. and Page B. (Eds.), "Computers support for environmental impact assessment," in IFIP, North-Holland, ISBN 0-444-81838-3, 1994.
47. Okubo T., Kubo K., Hosomi M., and Murakami A., "A knowledge-based decision support system for selecting small- scale wastewater treatment processes," Water Science Technol- ogy, vol. 30, no. 2, pp. 175–184, 1994.
48. Serra P., Lafuente J., Moreno R., de Prada C., and Poch M., "Development of a real-time expert system for wastewater treatment plants control," Control. Eng. Practice, vol. 1, no. 2, pp. 329–335, 1993.
49. Aarts R.J., Knowledge-based Systems for Bioprocesses, Tech- nical Research Centre of Finland, vol. 120, 1992.
50. Fox, M. S., & Smith, S. F. (1984). ISIS?a knowledge-based system for factory scheduling. Expert Systems, 1(1), 25–49.
51. Mar-Ortiz, J., Gracia, M. D., & Castillo-García, N. (2018). Challenges in the Design of Decision Support Systems for Port and Maritime Supply Chains. In Exploring Intelligent Decision Support Systems (pp. 49-71). Springer, Cham.

AUTHORS PROFILE

Altaf Alaoui, a Ph.D. at the Faculty of Sciences Kenitra, Ibn Tofail University, Morocco. Her research interests cover the study of data scientist, Machine Learning methods, clustering, estimation and prediction of the time series data.

Boris Olengoba Ibara, has over ten years of experience in image processing using Machine learning for classification and segmentation. He obtained a doctoral degree in the geomatics applied to environmental management.

His research interests cover the study of Knowledge Engineering, Geomatics, Image process, Machine Learning algorithms and ecology and environmental management

Badia Ettaki, is a professor at the Engineering School of Information Sciences in Rabat, Morocco. She obtained a doctoral degree in the Statistic Physic. His research interests cover the study of Content and Knowledge Engineering, Machine Learning algorithms and Big Data analytics.

Jamal Zerouaoui, is a professor at the Faculty of Sciences, Kenitra, Ibn Tofail University, Morocco. He received a doctoral degree in the physic science. Currently, his research interests cover modeling using Machine Learning methods in the frame of the physic area