Review of Data Mining Techniques in Environmental System

M.S. Chaudhari* & Dr. N.K. Choudhari**

*Head of Department, Computer Science & Engineering, Priyadarshini Bhagwati College of Engineering, RTM Nagpur University, Nagpur, Maharashtra, INDIA. E-Mail: manojchaudhary2{at}gmail{dot}com
**Principal, Priyadarshini Bhagwati College of Engineering, RTM Nagpur University, Nagpur, Maharashtra, INDIA.
E-Mail: drnitinchoudhari{at}gmail{dot}com

***Abstract*—**The emergence of various data mining algorithms and its application to various fields include medical imaging, network traffic analysis, environment system etc. Environment system now a day is the most important area of people's concern in today's world since it has daily impact on human beings life. May it be earthquake, soil erosion, deforesting, increasing summer temperature, rain fall density/intensity, flood occurrences and the most important is the impact of all these ES factors to directly and indirectly on the human beings and their behaviour. The capability of data mining algorithms of finding pattern in a data can be applied to Environment System data which is largely distributed, heterogeneous, sparse, multidimensional and heterogeneous. This paper gives a brief survey of essential steps, related algorithms and details processes that deals in designing and dealing with ES data that are essential in development of data mining tool for finding and interpreting patterns in environment system data set.

***Keywords*—**Clustering; Data Mining; Environmental System; Pre-Processing; Post Processing.

***Abbreviations*—**Environmental System (ES); Data Mining (DM); Knowledge Discovery from Data (KDD).

I. INTRODUCTION

FOR the decades the ES is largely ignored area for analysis purpose due to lack of proper data analysis tool or unavailability of the scientific tools. But the emergence of Data mining techniques and its wide use in different domains for finding or discovering patterns in large data sets have attracted Environmental scientist to consider this technique.

Data mining is the process of extracting hidden patterns from data and is becoming an increasingly important tool to transform this data into knowledge. It can be applied to data sets of any size (large volumes of data) and can be used to uncover hidden patterns to find valuable information but it cannot uncover patterns which are not already present in the data set.

Thus data mining is the overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms [Spate et al., 1].

II. LITERATURE SURVEY

2.1. The Process Detail

The detail processes involved in designing of DM Tool for environmental system are:

- *To Develop and understand the domain, to capture relevant prior knowledge and the goals of the end-user*

  Since domains of ES are very large, distributed and heterogeneous, we have to understand specific domain or field of study, develop the rough data set and goals of the end user.
- *To create the target data set by selecting a proper set of variables or data samples*

  The data generated by ES is tremendous and large which is not possible to use it entirely for analysis purpose. Hence we have to create target data set for analysis by using the observational samples over the years. Such data set can be large, unprocessed, scattered and distributed over the domain.
- *Data cleaning and preprocessing*

  Quality of data mining tool result is dependent on the quality of input data, and therefore the preprocessing step is crucial. The preprocessing is essential due to above qualities of data in the previous steps to obtain more clear, comprehensive and accurate results.
- *Data reduction and projection*

  In this step, depending on the problem, we try to simplify the set of variables (data set) of study domain to keep a relevant set of variables describing the system adequately and efficiently.

- *Choosing the data mining task, with reference to the goal of the KDD process*

  Different DM algorithms are in existence ranging from clustering to time series forecasting, classification, regression and many different techniques exist for different purposes depending on different requirements with specific domain. The relevant sets of algorithms are applied considering the domain of study.
- *Selecting the data mining algorithms*

  After selection of finalization of task and goals of the system, a set of methods needs to be chosen for searching patterns in the data.
- *Data mining*

  It is searching for patterns in data. Results from this stage will be significantly improved if previous steps were performed carefully.
- *Interpreting mined patterns*

  It is the interpretation of patterns obtained using above steps to get the specific results. This interpretation may consist of further iteration of previous steps.
- *Consolidating discovered knowledge*

  After interpreting found patterns, results are reported and documented, and /or used them inside the target system.

2.2. Essential Steps used in Design of Data Mining Tool for ES

To obtain more correct, accurate and useful results out of data mining tools of ES, following set of steps/algorithms must be applied on data set. All or some of the sequence of following main steps can be applied considering the domain of ES.

- Data cleaning & variable selection,
- Algorithm & parameters selection & application of algorithms,
- Interpretation of results.

Following are the different algorithms required in the process of data mining techniques or design of data mining tools. One or more algorithms are required to apply to get better results.

2.2.1. Data Cleaning & Variable Selection

- Pre-processing
- Data Reduction and Projection
- Visualization

2.2.2. Algorithm & Parameters Selection, Application of Algorithms

- Clustering and Density Analysis.
- Classification and Regression Method
- Association rule extraction/Analysis

2.2.3. Interpretation of Results

- Postprocesssing

Above algorithms are required:

- To better understand and preparation of data set.
- To detect imperfections in data sets and manage them in the proper way.
- To correctly accurately prepare data for the selected above DM algorithms [Gibert et al., 3].

III. DISCUSSION

3.1. Preprocessing

This step performs preprocessing operations like data cleaning, outlier detection (Outliers are objects with very extreme values in one or more variables), missing value treatment, transformation and creation of one or more Variables. This also includes removing the object completely (useful data may be discarded), replacing it with a mean or estimated value, duplicating the row once for each possible value if the variable is discrete, or excluding it from the analysis by modification of the data mining algorithm [Spate et al., 1].

Pre-processing step can be reduced to two main families of techniques:

- Detection Techniques

  In this technique imperfection in data sets or verification of the accomplishment of required assumptions for a particular analysis is detected.
- Transforming Techniques

  In this technique, transformations in the data set to correct the imperfections present data set for certain analysis technique is performed [Gibert et al., 3].

But due care should be taken to perform preprocessing since it may remove some useful data that may adversely affect the outcome in final step.

3.2. Data Reduction & Projection

When the number of variables is too high to deal with in a reasonable way, it may be convenient to apply a data reduction method. This kind of technique consists of finding some set with the minimum number of variables that captures the information contained in the original data set [Spate et al., 1].

Projection is accomplished by eliminating some variables totally or projecting the feature space of the original problem into a reduced fictitious space, with fewer dimensions [Spate et al., 1; Gibert et al., 3].

3.3. Visualization

A visualization technique plays an important role in the correct preprocessing of data. This is a powerful strategy for leveraging the visual orientation of sighted human beings.

One of the key points at which human interaction is often most fruitful is the visualization stages, during pre and post processing. The presence of outliers, missing values, errors, and unusual behavior are often first noted visually, enabling more detailed investigation later. Graphical methods should be the first stage of investigation for all datasets, even those whose dimension is too great to allow a comprehensive survey in this way [Gibert et al., 3]

3.4. Clustering & Density Estimation

A cluster is a collection of objects that are similar to each other and are dissimilar to objects in other clusters. Given a set of examples, the task of clustering is to partition these examples into subsets (clusters). The goal is to achieve high similarity between objects within individual clusters (interclass similarity) and low similarity between objects that belong to different clusters (intraclass similarity). Clustering can be viewed as a density estimation problem by assuming that the data was generated by a mixture of probability distributions, one for each cluster [Spate et al., 1]. It is concerned with grouping objects into classes of similar objects [Jindal & Bora, 4; Ester & Kriegel, 9].

Clustering group objects into classes with the objective of maximizing intra-class similarity and minimizing inter-class similarity (unsupervised learning) [Jindal & Taneja, 8; Dubes & Jain, 10]. It is unsupervised learning which find relevant features for categorization & clusters can be created by splitting complete data set into subsets or by unifying individual data items & consequently their representative [Spate et al., 5; Dzeroski, 6].

3.5. Classification & Regression Methods

In classification and regression, the identity of the target class is known a priori and the goal is to find those variables that best explain the value of this target, either for descriptive purposes (better understanding the nature of the system) or prediction of the class value of a new data point. Classical linear regression is a technique for finding the best linear equation defining the relationship between a numerical response variable and the independent variables, all of which should also be numerical [Spate et al., 1].

It is a two way technique (training and testing) which maps data into a predefined class [Jindal & Bora, 4; Spate et al., 5] and It discovers a model or function that maps objects into predefined classes (classification) or into suitable values (regression). The model/function is computed on a training set (supervised learning).Classification builds models for categorical classes and regression builds models for continuous classes [Jindal & Taneja, 8]. The tasks of classification and regression are concerned with predicting the value of one field from the values of other fields. The target field is called the class (dependent variable in statistical terminology). The other fields are called attributes (independent variables in statistical terminology). If the class is continuous, the task at hand is called regression. If the class is discrete (it has a finite set of nominal values), the task at hand is called classification. In both cases, a set of data is taken as input, and a model (a pattern or a set of patterns) is generated. This model can then be used to predict values of the class for new data [Chaudhuri & Dayal, 7].

Some classification techniques are neural networks, genetic algorithm, Fuzzy ARTMAP, Rough set classifier [Shrivastava & Shukla, 2].

3.6. Association Analysis

Association analysis is the process of discovering and processing interesting relations from a dataset [Spate et al., 1] and it discovers association rules [Ester & Kriegel, 9]. Association rules specify correlations between frequent item sets and identify specific relationships among data [Jindal & Bora, 4].

The association analysis is typically performed in two steps. First, all frequent item sets are found, where an item set is frequent if it appears in at least a given percentage s (called support) of all transactions. Next, association rules are found of the form $X \Rightarrow Y$, where X and Y are frequent item sets and confidence of the rule (the percentage of transactions containing X that also contain Y passes a threshold c [Chaudhuri & Dayal, 7]. In this case we would like to find any rules of the form $A \Rightarrow^{T} B$ that seem to occur in the data with frequency above a given threshold. Here A and B are just events of a certain type, with the rule if A occurs then B occurs within time T. A and B do not necessarily have to be the same variable [Spate et al., 5; Jindal & Taneja, 8].

Association rule mining finds interesting associations and/or correlation relationships among large sets of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. The market basket analysis used association rule mining in distributed environment. Association rule mining is used to find rules that will predict the occurrence of an item and based on the occurrences of other items in the transaction [Jindal & Taneja, 8], search patterns gave association rules where the support will be counted as the fraction of transaction that contains an item X and an item Y and confidence can be measured in a transaction the item i appear in transaction that also contains an item X [Ester & Kriegel, 9].

3.7. Post Processing

Apart from the important role of preprocessing, together with the correct selection of the data mining technique which will really answer the target questions, there is an important job to be done between getting the results of the data mining techniques and using them to support decision-making: to understand the results. Thus, it is important to:

- Identify the relevant information from the software outputs, depending on the aims of every particular analysis.
- Find the best way to present the selected results to the user in such a way that it becomes directly understandable, given that the final user does not know the technical details of the Data Mining method used [Gibert et al., 3]. Data mining techniques are important tools for knowledge acquisition phase of integrated model building, and because integrated models are very high in complexity, results are often correspondingly difficult to interpret and the decision maker may benefit from a post processing data mining step [Spate et al., 1].

IV. CONCLUSION & FUTURE SCOPE

Thus the design process of ES Tool using data mining techniques ranges from processing of rough data sets to transforming it into pattern for analysis. During this process several transformation on data sets has to be performed, algorithms have to be applied to get more finer and accurate result. As we go on applying more and more no of steps and algorithms, more correct would be the result. The above steps and algorithms can be applied to design tools for rainfall estimation, earthquake prediction etc.

REFERENCES

[1] J. Spate, K. Gibert, M. Sànchèz-Marr, E. Frank, J. Comas, I. Athanasiadis & R. Letcher (2006), "Data Mining as a Tool for Environmental Scientists", *International Environmental Modelling and Software Society*.

[2] P. Shrivastava & Dr. M. Shukla (2011), "A Brief Survey on Data Mining for Biological and Environmental Problems", *International Journal of Scientific & Engineering Research*, Vol. 4, Pp. 630–635.

[3] K. Gibert, J. Izquierdo, G. Holmes & M. Sànchez-Marrè (2008), "On the Role of Pre and Post-Processing in Environmental Data Mining", *Proceedings of International Environmental Modelling and Software Society (iEMSs)*, Pp. 1937–1958.

[4] R. Jindal & M.D. Bora (2011), "A Survey on Educational Data Mining and Research Trends", *International Journal of Database Management Systems* (IJDMS), Vol.5, No.3, Pp. 53–73.

[5] J.M. Spate, B.F Croke & A.J. Jakeman (2003), "Data Mining in Hydrology", *International Congress on Modelling and Simulation (MODSIM 2003)*, Editor: A. David, *Post, Modelling and Simulation Society of Australia and New Zealand Inc*., UWA UniPrint, Pp. 422–427.

[6] S. Dzeroski (2003), "Environmental Applications of Data Mining", *Lecture Notes on Knowledge Technologies*, University of Trento.

[7] S. Chaudhuri & U. Dayal (1997), "An Overview of Data Warehousing and OLAP Technology", *ACM Sigmod Record*.

[8] R. Jindal & S. Taneja (2010), "Comparative Study of Data Warehouse Design Approaches: A Survey", *International Journal of Database Management Systems (IJDMS)*, Vol. 4, No. 1, Pp. 33–45.

[9] M. Ester & H.P. Kriegel (1996), "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*.

[10] R. Dubes & A. Jain (1998), "Algorithms for Clustering Data", *Prentice Hall*.

Dr. N.K. Choudhari, Professor and Principal, Priyadarshini Bhagwati College of Engineering, Nagpur, India has a vast experience of 25 years in teaching and research field. He has guided several students and many of them awarded Ph.D. He has published numerous papers in various international journal and conferences on his specialized topic of data mining, signal processing and Non Destructive Techniques etc.

Manoj S Chaudhari is an Assistant Professor and currently working as Head of Department of Computer Science & Engineering, Priyadarshini Bhagwati College of Engineering, and Nagpur, India. He has total 12 years of teaching experience and guided many PG students. He has several international journal publications with data mining, theory of computation and natural language processing as area of interest.