ROUGH SET THEORY A PROMISING INSTRUMENT FOR DIAGNOSIS AND PREDICTION INCLUDING POLLUTION PHENOMENA

Madalina Arama, Margareta Nicolau, Gheorghe Batrinescu, Carol Lehr, Virgil Criste, Ana Anghel

National Research and Development Institute for Industrial Ecology – ECOIND, 71-73 Drumul Podu Dambovitei St., sector 6, 060652, Bucharest, Romania, phone: +40/21/4106716; fax: +40/21/4100575; 4120042; e-mail: ecoind@incdecoind.ro

The paper presents the use of Rough Set Theory (acronym RST) as a new emerging instrument to support the environmental decision in uncertainty conditions. New developed methodologies for environmental pollution diagnosis and prediction and their use in environmental impact/risk assessment are presented. Those methodologies can be successfully used to make prediction regarding pollution phenomena and seems to be a promising simple instrument to be implemented in order to adapt the measures to be taken in adequate time during incidental/accidental discharge so that pollution plume concentration (regardless the pollution type) to be estimated.

Keywords: RST, pollution prediction

1. Introduction

Rough Set Theory (English acronym RST) is emerging as a new instrument to support the environmental decision in uncertainty conditions. New type of methodologies for predicting the pollution concentration wave using Rough Set Theory is presented. Using this theory it is possible to establish:

-Dependencies that characterize the relation "pollution source" – "migration paths"- "pollutants";

-The attributes relevance for the purpose to predict the pollution concentration wave and the uncertainty measure in relation to the pollution phenomenon and its monitoring;

The Rough Set Theory is largely used in data-mining in the extraction and discovery of new connected data, in the data and information processing.

The introduction of new approximation models to the knowledge derivable from the quantitative and qualitative data represents a plus and conduct to the significant results in the fields such as decision support, decision analysis, artificial intelligence, finance, bio-informatics etc.

The pollution concentration prediction is a complex issue linked to the multitude of practical situations that a prediction model should synthesize for being of practical use, having in the same time the attribute of simplicity. The concentration prediction models of pollution plume are those that are conceived usually for pollutants and pollutants groups taking into account the physical and chemical properties, the discharge conditions. The paper is organized as follows: 1- Short background of Rough Set Theory, 2 – The methodological

steps to be used in prediction based on Rough Set Theory and 3 - Benefits from the use of such methodologies in pollution diagnosis and predictions.

2. Basics of Rough Set Theory

In order to emphasize RST capabilities for diagnosis and predictions, basics of this theory will be presented. The most important characteristics of any prediction is the uncertainty. Although it cannot be totally removed, it can be reduced. As the uncertainty is increasingly lower, the prediction is considered increasingly better. The decision making process based on prediction is considered decisions in uncertainty conditions when the occurrences of pollution consequences probabilities have a poor knowledge basis because extrapolations are made based on similar events, produced in similar conditions, modeled with similar models so uncertainty is unavoidable.

The most known quantification of the uncertainty is the probability. Uncertainty may come from the measurement of phenomenon and is known usually as measurement uncertainty and from the incomplete or lack of knowledge of the considered phenomenon usually known as epistemic uncertainty. If for the measurement uncertainty the statistic methods are able to successfully describe it, for epistemic uncertainty new instruments should be developed for facing it [1]. Among those new type of instruments those based on Rough Set Theory, Fuzzy Set Theory [2] try to complement the uncertainty quantification in a prediction process. Rough Set Theory represents a mathematic instrument with large capabilities in organizing knowledge when data of different quality (incomplete data, redundant data etc) are available. When numerous qualitative data are available, statistical methods can be successfully used but having less and non-homogeneous qualitative data makes the statistical data inappropriate. In those situations the proposed simple RST algorithm can be successfully used. The non-homogeneous data are selected and organized again in different ways so new hidden significance can be discovered.

Those capabilities to synthesize and reduce data so that possible rules to be discovered based on simple decision of type "if ".... "then" to be taken are valuable capabilities because they don't require special computational capabilities [3].

In RST all the data about the considered discussion topic are presented within a so-called information system (IS). In RST, a multitude of observations/events are presented having attached a certain amount of information within the table. The table rows are labeling the observations/events and the columns are labeling the attributes named as "A" that characterize the those observations/events.

An information system in RST is defined as:

$$IS= (U, A, V, d)$$

U= set of observations/events
A= set of attributes that characterize the observations/events
$V= \cup_{a \in A} V_a$ set of attributes' values

d: U x A$\rightarrow$V is a function defined on U x A taking values from V set, the set of all attribute values that shows that any observation/event from U has a value from V for any attribute from A.

The attributes set A is formed from the union of two disjoint sets namely the conditions attributes C and the decision attributes D. In order to use information from this information system and to analyze, select and reduce the available information based on its utility in order to discover if it offers or not useful information (that means for example that information is not repeating) the theory introduces the indiscernible relation among two observations/events having the same characteristics. For any sub-sets of type $A_n \subseteq A$ considering that index "n " takes values from 1 ÷ N (N being finite) and representing A_1 – attributes of type 1, A_2 - attributes of type 2, ... A_n attributes of type n, the indiscernibility relation is introduce as:

$$IND\ (A_n) = \{(x, y) \mid \forall a_n \in A_n \rightarrow \mathbf{d}(x, a_n) = \mathbf{d}(y, a_n), x, y \in U\}$$

For any attribute of index "n" from the sets A_n the values of observations/events x and y from U (the universe of discourse) that are indiscernible are the same. With this indispensability relation introduced for different sub-sets of attributes, the universe of observations/events describing the topic can be partitioned in objects/events indiscernible events. A generic elementary set E_{An} is formed from indiscernible objects/events one from another in relation with any subset of attributes indexed by any "n" from sets A_n i.e. $E_{An} = IND\ (A_n)$.

If $A_n \subseteq A$ and $A_n = A$, then U/A represents the finest granularity of the knowledge because contains the greatest number of attributes indexed "n" that represents all the attributes known at certain moment. U/$\varnothing$ represents lower knowledge granularity (the lack of detected observations/events knowledge).

Strictly speaking the universe of discourse U formed by observations/events that are of interest for the corresponding topic should be infinite in the real world, but for the most cases we don't have access to the their multitude [4] so that we are forced to limit to a finite number accessible to us at a certain moment.

The observations/events can be indiscernible in relation with the subset of condition attributes C or decision attributes D, or in relation with both A=C$\cup$D. The elementary sets generated by decision attributes are named concepts (classes) [4].

The information system described by RST through the proposed algorithm realizes a classification of observations/events in classes based on the characterization of attributes' values. For the classification consistence, the observations/events that have the same attributes' conditions should to be in the same class. One can induces decision rules based on available known/measured values for the characteristics(attributes) for the considered observations/events at a certain moment. One special capability of RST is that it can deal with clear cut concepts as well as with vague concepts using mathematical construction named rough set. A vague concept is described by objects/events that, for the same condition attributes' values, have different decision attributes values. Such a vague concept described by a rough set

$X \subseteq U$, can be replaced according to RST with two classical sets: the lower approximation set $A_*(X)$ - having objects/events that for sure are part of the concept and $A^*(X)$ – having objects/events that possible are part of the concept and the upper approximation set of the rough set X considering the subsets of attributes from A . $R_B(X)= A^*(X) - A_*(X)$ is called region of boundary of X and is determined by the attributes' subsets from A. A rough set has always a non-empty boundary region.

3. Environmental pollution diagnosis and prediction based on RST and its use as environmental impact/risk assessment new methodologies

The literature about the use of RST in diagnosis is growing [5],[6],[7],[8],[9],[10],[11],[12] and it is based on the fact that any system that is monitored represents a data base with measurements values for different parameters characterizing the functionality of the corresponding system. From those databases using RST as a data-mining tool different regularities/patterns [13] can be discovered so that they can be used to take decision related to the regulation of the system operation. All the information available with reference to the values of parameters (attributes) set that characterize the system elements at the certain moment can be used to assess the operational status of that system to have the possibility to predict future dysfunctions in the normal operation or even the system failure [14].

When for the set of observations/events characterized by the parameters describing the normal or abnormal/failure of system operation there are parameters values at the lower limit of normal functioning, functionality might become uncertain. Is the situation when for 50% of relevant parameters characterizing the system operation the values are normal while for the other 50 % of the relevant parameters are abnormal. Those should be situations that brings uncertainty in judging those situations. Using RST capabilities in managing uncertainty there is the possibility to derive valid rules for decisions.

Such situations are not rare in the environmental pollution situations and that is precisely why using all available data the expert should be able to judge the extent of the pollution and the risk of different kind of pollution consequences following a hazardous event that is considered the pollution, leading finally to a significant environmental impact for one environmental segment such as air, water, soil separately or for more environmental compartments in the same time.

Judging those kind of situations means actually making a good diagnosis in relation to significance of the environmental impact and assessing the environmental risk by considering the probabilities of occurrence of a sum of consequences following the corresponding diagnosed significant environmental impact on one or more environmental compartments.

Because RST is able to face both measurement and epistemic uncertainties it is especially suited for environmental diagnosis and predictions.

The steps of such type of methodologies are briefly presented in Fig. 1. Step 1 can be conceived for 1 or more compartments RST supporting a separate or an integrated approach for the environmental impact/risk assessment. It is the most important step because it establishes the conceptual

model suitable for the corresponding assessment situations. Using RST data of different quality can be used and synthesized so that finally they can be processed and after reducts (sets with minimal but relevant information) computation the decision rules to be obtained and to support the decisions in uncertainty condition, uncertainty that RST manages by a an approximation process.

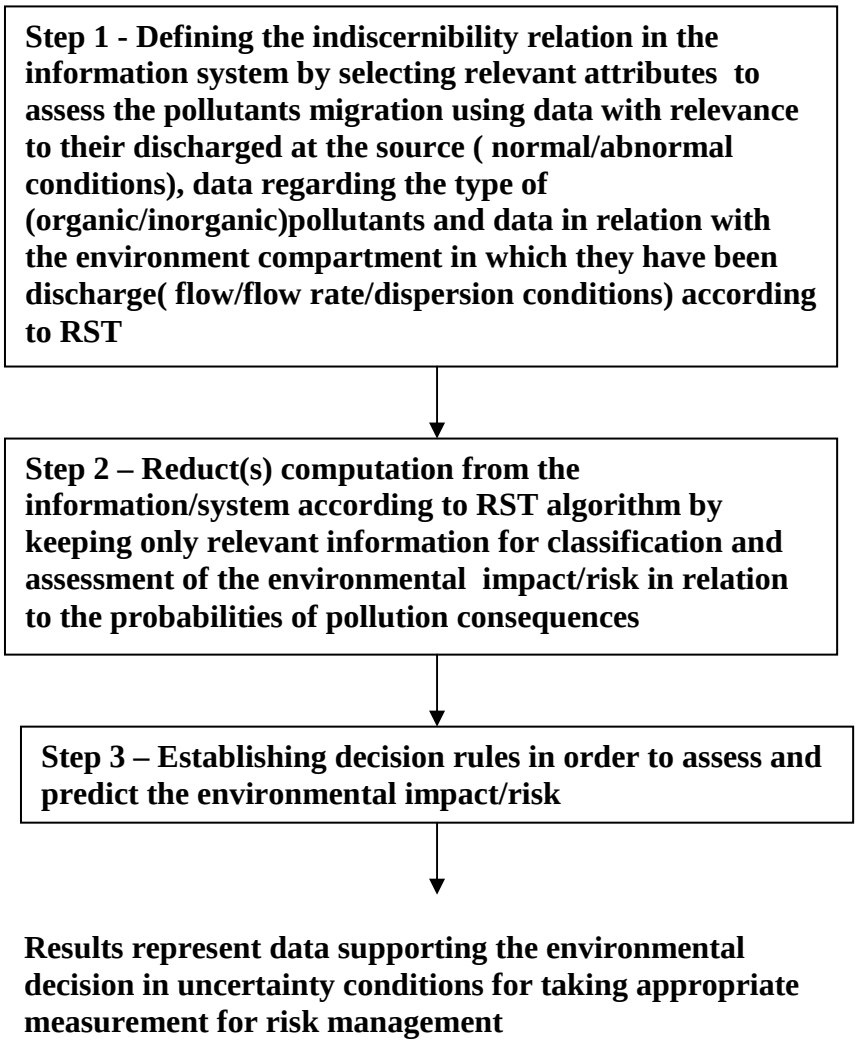

Fig. 1. Methodological steps for environmental impact/risk assessment using RST

Those types of methodological endeavors using RST are based essentially on one or more reference reducts (decision table with consistent information) that can be used to characterize pollution in order to detect trends in future pollution and to take appropriate measures for the pollution control. All those methodologies use as basic definition for risk concept the probability that certain specific effects/certain consequences can be produced following the occurrence of a hazardous event such is a significant pollution in a defined period of time and defined circumstances.

4. Benefits from the use of such methodologies in pollution diagnosis and predictions

RST brings a number of benefits for the environmental risk diagnosis and prediction because:

1. it offers a tool to organize non-homogeneous information in relation to the environmental states/environmental observations;
2. it can present the information in an aggregated forms (the attributes can be formed by simple or aggregated indicators)
3. it allows to express the quality of classification by presenting the accuracy of classification as well as the degree of consistency measures;
4. it can be used to show the trends of increasing, maintaining or decreasing of pollution at a certain moment in time at certain location, with reference to the legal allowable limits of pollutants concentrations that can generate significant impact. These trends are used to emphasize the possibility of occurrence of hazardous events at different environmental targets. This way the targets' risk can be characterized. The probabilities occurrence and evolution of different severity consequences within a specified period of time in specified circumstances can be estimated.

5. Conclusions

Those methodologies can be successfully used to make prediction regarding pollution phenomena and seems to be a promising simple instrument to be implemented in order to adapt the measures to be taken in useful time both during normal and incidental/accidental discharges so that pollution plume concentration (regardless the pollution type) to be estimated. Extension of using those type of methodology in the field of the environmental impact/risk assessment should improve the actual extensively descriptive methodology studies. It complements those studies with a set of criteria and decision rules so that the conclusions can be reviewed in the light of new evidences those new information being incorporated so that adequate decisions to be taken.

References

[1] BERNARDINI, A. AND TONON, F., "*Aggregation of Evidence from Random and Fuzzy Sets*" – Zamm-Z. Agnew. Math. Mech. 84.10-11, (2004) 700-709. Web. 29 September 2009 <www.zamm-journal.org>

[2] ZADEH, L.A.*; "Toward a Generalized Theory of Uncertainty (GTU) – an Outline*". IEEE International Conference on Granular Computing, on Volume 1, (25-27 July 2005): 16. web.29. September, 2009

[3] MAGRO, C. M., PINCETI, P. *"A confirmation technique for maintenance using the Rough Set Theory"*. Computers & Industrial Engineering. 56(2009) 1319-1327.

[4] YIYU YAO; *"Probabilistic rough set approximations".* International Journal of Approximate Reasoning 49 (2008) 255-271

[5] LI ZHIYAO, WANG MOYU, MA XINKE, SHEN XIAOLIU "*High Risk Management Model For The Power Enterprise Based on Rough Set Theory*" System Engineering Procedia 3 (2012) 63-68.

[6] ARMAGHAN ABED-ELMDOUST, REZA KERACHIAN "*Wave height prediction using the rough set theory*" Ocean Engineering 54 (2012) 244-250.

[7] P. PATTARAINTAKORN, N. CERCONE, K NARUEDOMKUL "*Rule learning: Ordinal prediction based on rough sets and soft-computing*" Applied Mathematics Letters 19 (2006) 1300-1307.

[8] DOMINIK SLEZAK, WOJCIECH ZIARKO "*The investigation of the Bayesian rough set model*" International Journal of Approximate Reasoning 40 (2005) 81-91.

[9] A.ZABEO, L.PIZZOL, P.AGOSTINI, A. CRITTO, S.GIOVE, A. MARCOMINI "*Regional risk assessment for contaminated sites Part 1: Vulnerability assessment by multicriteria decision analysis*" Environment International 37 (2011) 1295-1306.

[10] R.M. Darbra, E. Eljarrat, D. Barcelo "*How to measure uncertainties in environmental risk assessment*" Trends in Analytical Chemistry, Vol. 27, No. 4. 2008.

[11] Linyu Xu, Guiyou Liu " *The study of a method of regional environmental risk assessment*" Journal of Environmental Management 90 (2009) 3290-3296

[12] KEJIANG ZHANG, YUANSHENG PEI, CHANGJING LIN "*An investigation of correlations between different environmental assessments and risk assessment*" Procedia Environmental Science 2 (2010) 643-649.

[13] KIYOSHI HASEGAWA, MICHIO KOYAMA, MASAMOTO ARAKAWA, KIMITO FUNATSU *"Application of data mining to quantitative structure-activity relationship using rough set theory"* – Chemometrics and Intelligent Laboratory Systems.

[14] IFTIKHAR U. SIKDER, TOSHINORI MUNAKATA "*Application of rough set and decision tree for characterization of premonitory factors of low seismic activity*" Expert Systems with Application 36 (2009) 102-110.