IJARESM
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING, SCIENCE & MANAGEMENT

MATHEMATICAL MODELLING TO SOLVE LIMITATION OF DATA MINING TECHNIUES FOR AIR POLLUTION PROBLEM

Ms. BhavikaM.Tailor[1], Mr. Rasik R.Shah[2]

Assistant Professor, CGPIT, UTU, Bardoli, Gujarat, India1

Assistant Professor, SNPIT & RC, Umrakh, Bardoli, Gujarat, India2

Abstract: ***Data mining techniques extract hidden information from large amount of data but this process is complicated and difficult as well as some time unable to provide meaningful information. Prediction involves some fields in the data set to predict unknown variables of interest while description focuses on finding patterns describing the data that can be interpreted by humans. Inthis analysis toll important thing is to build mining model and it is one of the large process that includes everything from defining the basic problem that the model will solve, to deploying the model into a working environment. In this paper researches discussed some of the data mining limitation and their solution with Mathematical and Statistical tools for solution of environmental problems.***

Key Words: Data Mining, Faulty Data, Air Pollution, Stability Analysis

I. INTRODUCTION

Safety of the environment has to be an important part of sustainable broad growth of any country. Healthy environment development is especially important for developing country when awareness of the risks of environmental degradation has enlarged deeply. In 21st centuries the environmental problems in India are increasing speedily. The rising economic progress and a speedily growing population that has taken the country from 300 million citizens in 1947 to more than one billion citizens at present is putting a damage on the environment, infrastructure, and the countries natural assets [1]. To solve these problems computation techniques are so valuable. Among all these tools and techniques Data Mining is more effectual to solve special types of environmental problems.

Data mining is the procedure of finding patterns from data. Data mining is becoming a gradually more significant tool to convert these data into information. There are several major *datamining techniques* have been developing and using in data mining projects recently including *association*, classification, clustering, prediction, *sequentialpatterns* and *decisiontree [2]*. The selection of methods for data analysis modeling and algorithms depends on quality and quantity of data but all these tools and Data Mining techniques have some limitations.

II.DATA MINING FOR SOLUTION OF ENVIRONMENT PROBLEM

In air pollution control and management, the cluster analysis of collected air pollutants can recognize and the spatial variation with time, so that the administration can adopt appropriate policies for environment protection. Several factors from input data, such as the data length, scale, transform, etc., can cause different analysis results. The focal aim of many environmental system analyses is to maintain posterior decision making to recover either management or control of the system. Intelligent Environmental Decision Support Systems (IEDSSs) are amid the most promising approaches in this field. IEDSS are included models that give domain information by means of analytical decision models, and permit access to databases and knowledge bases to the decision maker [4, 5].

III.LIMITATIONS OF DATA MINING FOR ENVIRONMENTAL PROBLEM

As discussed above data mining extremely powerful tools but today in the market it has not self-satisfactory applications. The first reason behind that is data mining will give meaningful information from data but if data is faulty then it is difficult to find consequential information from the data. The second reason is when two or more than two patterns are equally probable then which one is batter will not provide better. Third reasons some time data exploration and modeling are time consuming and so on. As far as Environmental data is concern it is future prediction is difficult[2, 6].

IV.MATHEMATICAL MODELLING FOR SOLUTION OF DATA MINING LIMITATIONS

In Data mining techniques data cleaning steps are crucial, up till now whatever data mining technique algorithm invent are not up to that level so that it is easy to find faulty data from the field.For example suppose temperature field contain temperature data whose average is suppose 30 but one of its observation is 350 which is faulty so it is necessary to remove such observation from the data but data mining technique unable to do such kind of things so researchers suggest following technique for find out the faulty data. This method is useful to determine whether data values should be rejected is also called as faulty data eliminated method.

A. Faulty Data Eliminated Method.

1. Calculate the mean and the sample standard deviation of the complete data set.
2. Obtain R corresponding to the number of measurements taken from R table [7]. Assume the case of one doubtful observation first, even if there appears to be more than one.
3. Calculate the maximum allowable deviation: $|\mathbf{x_i} - \mathbf{x_m}|_{\mathbf{max}}$.
4. For any suspicious data measurements, obtain $|\mathbf{x_i} - \mathbf{x_m}|$.
5. Eliminate the suspicious measurements if:$|\mathbf{x_i} - \mathbf{x_m}| > |\mathbf{x_i} - \mathbf{x_m}|_{\mathbf{max}}$.
6. If these results in the rejection of one measurement, assume the case of two doubtful observations, keeping the original values of the mean and standard deviation, and of the original number measurements. Go to step 8.
7. If more than one measurement is rejected in the above test, assume the next highest values of doubtful observations. For example, if two measurements are rejected in step 5, assume the case of three doubtful observations, keeping theoriginal values of

the mean, standard deviation, and the original number of measurements as the process is continued.

8. Repeat the above calculations (steps 2 – 5) sequentially increasing the number of doubtful measurement possibilities, until no more data measurements need to be eliminated.
9. Now obtain the new value of the mean and sample standard deviation of the reduced data set [7].

This procedure can easily find faulty data from the whole data set and will give normal data for decision which is highly beneficial for solution of Environmental Problem. The procedure to find stability of pattern involved with different components. It is difficult to find stability or fluctuation from each and every component involved in the air pollution problem for that researchers find stability of pattern form its pattern. If in pattern more than one component involved then it is necessary to find combine stability or fluctuation forcomponents which will give us fluctuation or stability pattern.

B .Stability Measurement Model: The basic framework of the procedure to find stability and uncertainty is presented in the figure 1 given below.

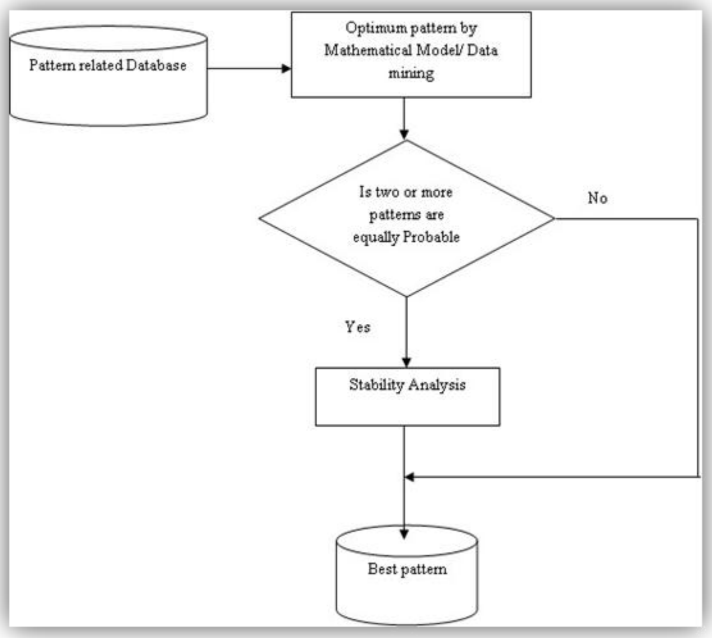
Pattern related Database
Optimum pattern by Mathematical Model/ Data mining
Is two or more patterns are equally Probable
No
Yes
Stability Analysis
Best pattern

Figure 1: Stability measurement model

Stability analysis Procedure from optimum pattern:May be possible optimum pattern contains one or more than one component so it is necessary to find fluctuation from each component from the past data and then find combined fluctuation for all the pollutants which are involved in pattern. If the procedure is applied for system form than it is described as follows

System Modeling**:** Suppose for any optimum pattern, Z as a function of the n measurable crops x_i, i = 1,2,3…N

$$z = y\,(x_1, x_2, x_3, x_4, x_5, x_6\ldots, x_n) \qquad (1)$$

Where,x_i = i^{th}componentinvolved in pattern. Equation (1) is called the system equation for the measurement. The component involved in this system may be inter correlated or may not inter correlated so according to that there are two cases.

Case-I Components are Independent*:* Suppose all components involved in optimum pattern are independent from each other. In this case it is necessary to find fluctuation form each and every component and then find combine fluctuation for allcomponents. For any individual componentsof optimum pattern, suppose n different measurementvalues areX_1, X_2, ... ,X_n, then the random fluctuation in a measurement X is estimated by standard deviation

$$(u_m) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n\,(n-1)}} \qquad (2)$$

Equation (2) will give us fluctuation of one measurement, which is involved in optimum pattern. If more than one component were involved in optimum pattern then general of fluctuation or standard deviation formula is given as follows

$$(u_m)_k = \sum_{k=1}^{N}\left(\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n_i(n_i-1)}}\right) \qquad (3)$$

For first summation k=1,2,3,…….N and for second summation i=1,2,3,….n

Equation (3) will give standard deviation for all the k components which are involved in optimum pattern. From these individual fluctuations, combine fluctuations as well as combine mean for optimum pattern can be derived as follows

$$\sigma = \sqrt{\frac{\sum_{k=1}^{n} \mathrm{n_k}\,(\mathrm{u_m}^2)_k + \sum_{k=1}^{n} \mathrm{n_k}\,(\mathrm{d_{k2}}^2)}{\mathrm{n_1 + n_2 + n_3 + \cdots + n_k}}} \qquad (4)$$

Where $d_k = |\,\overline{X_k} - \overline{X_{123\ldots k}}\,|$ and

$$\bar{X}_{123\ldots k} = \frac{\sum_{k=1}^{n} n_k X_k}{\sum_{k=1}^{n} n_k} \qquad (5)$$

k= 1,2,3…n

Equation (4) and Equation (5) will give us combine standard deviation and combine mean of components which are involved in optimum pattern. This combine standard deviation and combine mean are useful for find stability of optimum pattern.

Case-II Components are Inter-correlated: In above cases optimum patterns components are not correlated but if they are inter-correlated then it is necessary to find inter-correlated fluctuation or deviation for stability analysis. For any component of optimum pattern n measurements yields the profit values X_1,X_2, ... ,X_n, then the random uncertainty in a measurement X is estimated by the standard deviation

$$u_{ran} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2}\,. \qquad (6)$$

Where, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

The combine fluctuation for all the components, which are involved in optimum pattern, is given by the formula

$$\sigma_y^2 \cong \sum_{i=1}^{n} c_i^2 \sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} c_i c_j \operatorname{cov}(x_i, x_j) . \qquad (7)$$

If two components are correlated than this correlation is quantified by a correlation coefficient. The correlation coefficient can be either estimated heuristically or computed from statistical samples. Correlation coefficient ρ_{ij} for two crops x_i and x_j is defined as

$$\rho_{ij} = \frac{\operatorname{cov}(x_i, x_j)}{\sigma_i \sigma_j} , \qquad (8)$$

Where, σ_i and σ_j are the standard deviations in x_i and x_j. Using this definition, we can rewrite Equation. (6) as

$$\sigma_y^2 \cong \sum_{i=1}^{n} c_i^2 \sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} c_i c_j \rho_{ij} \sigma_i \sigma_j . \qquad (9)$$

The statistic that quantifies this spread is the standard deviation. The standard deviation is just the square root of the variance

$$\sigma_y \cong \sqrt{\sum_{i=1}^{n} c_i^2 \sigma_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} c_i c_j \rho_{ij} \sigma_i \sigma_j} . \qquad (10)$$

This standard deviation of the components is useful to determine more stable or less risky optimum pattern for improvement of data mining technique if these two techniques include then it is highly beneficial for practical application of data mining.This procedure again if includes in data mining techniques it will short out the limitation of mining and provide batter solution for environmental problem.

V. CASE STUDY

The environmental problems are happened because of doubtful aspects and pollutants. The doubtful measurements of pollutions are highly effected on the results. In data mining techniques such kind of doubtful data play significant role to make wrong.

A. FaultyData EliminatedMethod for Environmental Problem

The faulty data eliminated method is described here in details with environmental data. The algorithm for faulty data correction is discussion in article apply at here for environmental data specifically for air pollution of metro cities. Researchers have taken three cities data Delhi, Kolkata and Vadodara respectively which contains SO_2, NO_2 and RSPM pollutants data of year 2003 among so many pollutants. The processes apply for Delhi, Kolkata and Vadodara for the pollutant $SO_{2,}$ NO_2 and RSPM then the summaries result is as follows in table 1 contain:

TABLE I: LIST OF FAULTY DATA WHICH ARE ELIMINATED

City	**Delhi**			**Kolkata**			**Vadodara**		
Pollutants	SO_2	NO_2	RSPM	SO_2	NO_2	RSPM	SO_2	NO_2	RSPM
Eliminated Data	50	197	670	35	200	338	55	68	256
	49				150		37		199

After this procedure the remaining database is normal which will useful to give an accurate decision for future environmental problem. So data mining technique is specified to reduce some air pollution from environment. This is the small contribution of researcher to give precise data mining techniques.

B.Consistency Analysis Procedure from Optimum Pattern

The researchersare taken an assumption that all three cities given in table 1 have equal pollution pattern and thereafter try to find which city is more consistency polluted. The answer of this question is possible by the developed data mining technique or constancy measurement model which is easily find the required output. The pollutants SO_2, NO_2 and RSPM may be inter-correlated or may not inter-correlated so according to that there are two cases. Here is a calculation of Delhi pollutants.

Case-I Pollutants in the pattern are Independent: Suppose that the pollutants in this pattern are independent. So the researchers have found the fluctuation form each and every pollutant SO_2, NO_2 and RSPM, and then find the combine fluctuation of all pollutants. Here thirty five measurements of pollutants are taken for experiment. The combine fluctuation is obtained by combine standard deviation. For that the necessary discussion is below. So when the pollutants are not inter-correlated the combine variance of Delhi, Kolkata and Vadodara pollutants are 23307.25296, 5076.24491 and 1391.985372 respectively. So the conclusion is that the combine variance of Vadodara pollutants is less than other two cities pollutants. So the Vadodara city is less consistency polluted than other two cities.

Case-II Parameters in the patterns are Inter-correlated: Assume that the pollutants SO2, NO2 and RSPM are inter correlated then it is necessary to find inter correlation between them according to our techniques. The Summary of this procedure is described in table as follows.

TABLE II: SUMMARIZE TABLE FOR CONSTANCY MEASUREMENT MODEL

	Delhi	Kolkata	Vadodara
Correlation coefficient	$\rho_{12} = 0.106218$ $\rho_{23} = 0.222386$ $\rho_{13} = -0.13891$	$\rho_{12} = 0.441958$ $\rho_{23} = 0.301411$ $\rho_{13} = 0.271006$	$\rho_{12} = -0.0496$ $\rho_{23} = 0.102641$ $\rho_{13} = -0.22674$
Combine Variance	$\sigma_y^2 = 48534.92107$	$\sigma_y^2 = 26139.0759$	$\sigma_y^2 = 7472.777783$
Combine Standard Deviation	$\sigma_y = 220.366425$	$\sigma_y = 161.575835$	$\sigma_y = 86.445229$

From the Table 2 It is clear that when all pollutants SO_2, NO_2 and RSPM are inter-correlated then the combine variance of Delhi, Kolkata and Vadodara is 48534.92107, 26139.0759 and 7472.777783 respectively. So the outcome is that when the pollutants described in table 1 are inter-correlated then the combine variance of Vadodara is less than other two cities. It means the pollution is fewer consistence in Vadodara than other two cities.

VI.CONCLUSION

For real world application of data mining improvement in techniques and algorithm is required so that software become effective and significant which could be useful to trust the people of other sectors in software accuracy.

REFERENCE

[01] The National Eleventh Five – year Plan for Environmental Protection (2006 – 2010)

[02] Jeffrey W. Seifert "CRS Report for Congress Data Mining An Overview"RL 31798 December 16, 2004

[03] Sheng-Tun Li and Shih-Wei Chou, "Multi-Resolution Spatio-temporal Data Mining forthe Study of Air Pollutant Regionalizationl" Proceedings of the 33rd Hawaii International Conference on System Sciences - 2000

[04] Jessica Spate , Karina Gibert, MiquelS`anchezMarr`e, Eibe Frank, Joaquim Comas, IoannisAthanasia, Rebecca Letcher, "Data Mining as a Tool for Environmental Scientists"Australian National University, 2002..

[05] U.W.Tang,"Capture and Data Mining of Urban Air Pollution" Department of Civil and Environmental Engineering University of Macau,Macau (PR China),2007

[06] Jayce.Jackson, " Data Mining: A Conceptual Overview" Communications of the Association for Information Systems.Page no 267-296, Volume 8, 2002;

[07] Shephen M Ross, "Pierce Criterion for the Elimination of Suspect Experimental Data" Journal Of Engineering Technology, Fall 2003