AN EFFICIENT FRAMEWORK OF PREDICTIVE DATA MINING UNDER A CASE STUDY OF GAS ENERGY PRODUCTION IN PAKISTAN

Sehresh Khan[1], Maqbool Uddin Shaikh[2], and Altaf Khan[3]

[1]Department of Computer Science, COMSATS Institute of Information Technology, Islamabad, Pakistan
sehreshkhan@comsats.edu.pk
[2] Department of Computer Science, Preston University, Islamabad, Pakistan
maqboolshaikh@preston.edu.pk
[3]Department of Mechanical Engineering, University of Engineering and Technology, Taxila, Pakistan
altaf_khan768@yahoo.com

Abstract

A lot of research has been carried out to study the energy crisis in Pakistan by using the predictive data mining techniques. Many researchers have tried to analyse the situation by using different frameworks but unfortunately the authors of this paper did not find any complete, cost effective and efficient framework in the literature. We therefore proposed a framework for developing a predictive data mining system, which will efficiently work in scenario of forecasting with an efficient flow of work. In this conceptual framework authors have showed twelve steps and their sequence of working in the form of algorithm. We applied the proposed framework on the case study for prediction of natural gas energy production in Pakistan to give a solution of the current energy crises. In this case study efforts were made to collect the historical data by covering different geographical areas of Pakistan and circumstances. Authors designed an "Energy Analytical Data Mart" to store the historical and continuously growing data. In this case study we have presented two approaches of forecasting the level of natural gas energy production in Pakistan using the artificial intelligence field neural network.

Keywords - Data Mining, Energy Crises, Framework, Gas Production Forecasting, Predictive Data Mining.

1 INTRODUCTION

Energy is an important input contributing to the economic growth and development. The energy security and its optimized usage are as important as geographical and economic security of any country. Although Pakistan is one of the countries, which are blessed with natural resources, yet it is dependent on the external resources to fulfil its Energy needs. This is because the internal natural resources of Energy are not managed efficiently and used effectively. Other than Natural Gas, there are other predictable energy resources such as solar photovoltaic, solar thermal power, wind and coal energies, which can be used to overcome the Natural Gas energy crises in Pakistan.

A large number of countries in the world despite being rich in natural energy resources are unable to use these resources in cost-effective way due to unavailability of human experts and technological resources. In 2009, Nayyer, and Zeeshan [1] raised the view point to rely on other energy resources e.g. geothermal energy besides natural Gas, because one day natural resources will be consumed from all over the world.

In previous years the demand of energy has been observed increasing drastically with the advancement in technology. In 2008 Tahir reported that energy demands in Pakistan will be doubled in next few years and the current demand may swell to seven times in year 2030 [2]. According to Munawar A. Sheikh [3] conventional energy resources such as gas, oil, coal etc. of Pakistan were the 99% of the total energy supply in the country in 2007 [3]. Keeping in mind the factors of increase in the demand of energy, availability of natural resources and current situations of energy resources in the country, described by [1], [2], [3] it is very important to analyze future energy production and usage

requirements to deal with the upcoming challenges in the field of energy. In our research work we have focused on the data of "Natural Gas Energy" which is the prime factor to be considered to address energy crises in the country. Research on other energy sources will be done later after successful research on predictive data mining for the Gas Energy.

The fields of data mining have considerable importance in extracting novel and actionable information from large databases [4] and also provide techniques of prediction and analysis of data. P. S. Bradley et. al. [5] discussed some factors to show the importance of data mining. From last few years data mining techniques are applied in various fields that fabricate productive and efficient decisions. These techniques are helping to decision makers to analyze the huge databases to extract novel patterns such as decision tree, clustering, prediction modeling, classification, pattern classification etc. [6], [7], [8], [9], [10], [11], [12], [13].

For real time domains there are different evolutionary models of data mining which worked successfully. Munir et. al. [14] used the combination of evolutionary algorithms and some machine learning methods in the field of eco-informatics and produced results with very low percentage of errors. Fangwen et. al. [15] worked on the hybrid approaches for the forecasting model research on stock price data mining.

In the literature we find only few frameworks, which relates to the field of data mining or knowledge discovery. As Usama et. al. [15] presented a framework for the discovery of knowledge in the databases. They discussed the links between the fields of KDD, data mining and other related fields with their applications and challenges [16].

In 2005 Rie and Tong [17] presented a framework, which helps in learning the structure of forecasting for the unlabeled data. In their paper they presented a novel approach for semi-supervised learning. Their framework integrated the algorithms, which formulates the structural learning problems theoretically.

Mitra and pal [18] presented a survey of IEEE transactions on neural networks and come up with a soft computing framework of data mining. In their research they analyzed the current techniques for data mining e.g. neural networks, genetic algorithms, fuzzy logics etc.

After reviewing the literature of data mining frameworks, authors of this research come to the conclusion that a detailed and efficient framework for the predictive data mining is the current need of this field. Authors of this research have proposed a framework for developing a predictive data mining system, which can efficiently work in any scenario of forecasting with an efficient flow of work rather only selecting an appropriate algorithm. In this conceptual framework authors have showed ten required steps which also applied on the case study of prediction of gas production in Pakistan.

In this paper authors are presenting the problem definition in section 2, the proposed conceptual framework in section 3. The elaboration of all the process of the framework is presented in the sub sections of section 3. The conclusion, issues and future work are presented in section 4.

2 PROBLEM DEFINITION

In the literature there are several solutions presented in the history of successful data mining for forecasting energy production in Pakistan. Different authors tried to solve the problems by using the historical data and gain the optimal forecasted novel facts of different fields. However, still not much work has done to develop a framework which can present a complete, detailed description of their processes which can give much more efficient analysis of any forecasting scenario with the help of artificial intelligent agents. In this research paper authors have presented a conceptual framework which can show step by step processes of the presented framework. The whole processes are described for successful, efficient, and accurate forecast by using the available historical data of gas production

3 PROPOSED SYSTEM

The Proposed framework having twelve processes is graphically presented in fig.3.1. The detailed description of all processes with the example of the Gas Production forecasting problem is presented in the sub sections presented below.

In the Fig.3.2 working of the framework is presented in the form of a procedure. In the given procedure T is the threshold for checking accuracy of the developed mining model, AlgoS is the selected algorithm from already available algorithms in the literature, AlgoR is the required algorithm as per the solution requirements gathered in step.2-a, AlgoU is the updated algorithms, Acc represent accuracy of the developed mining model, AlgoN is the new created algorithm.

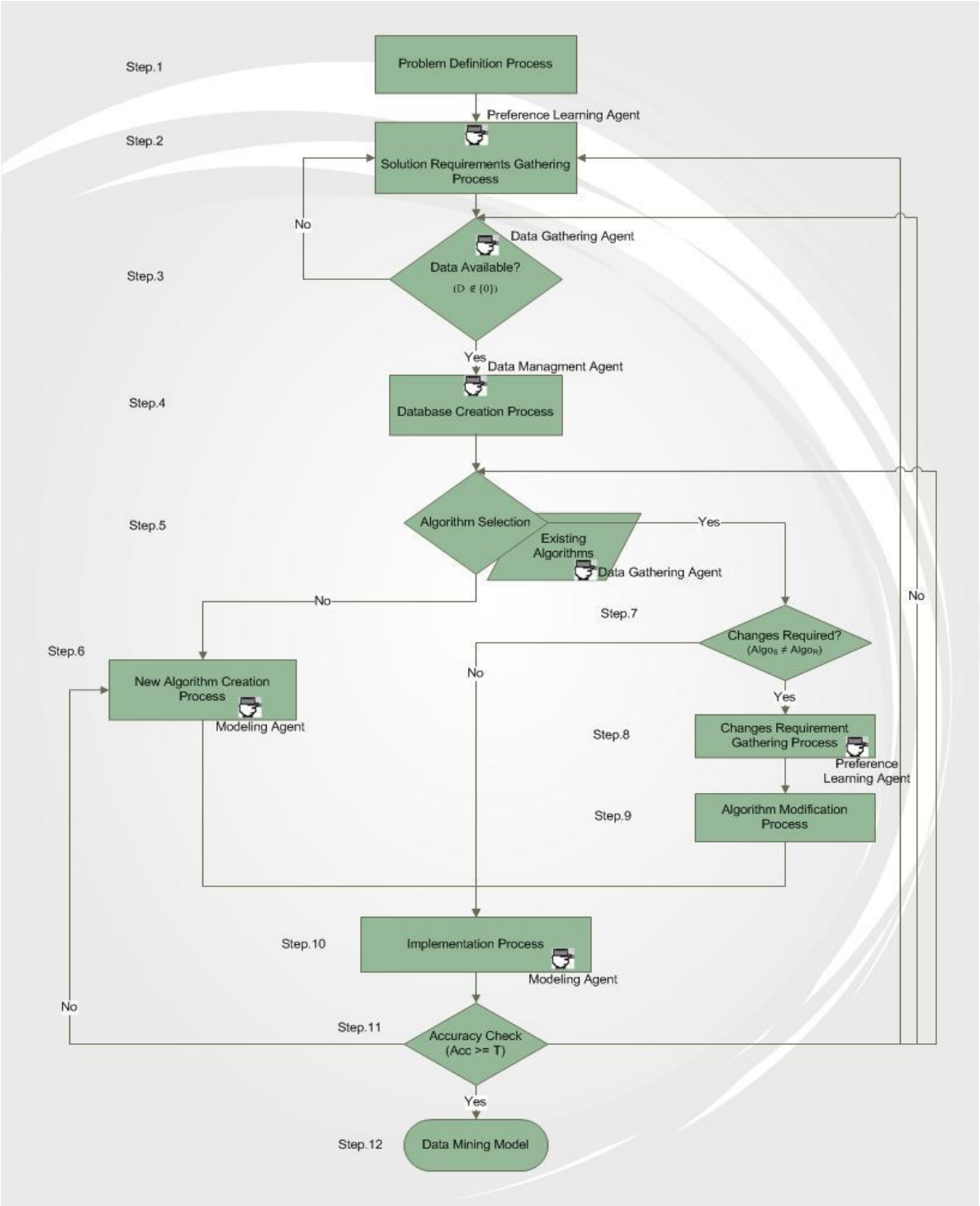

Fig.3.1. Proposed frameworks for development of forecasting model

Keys: Rectangle shows process, Diamond represents decision making process, Parallelograms represent data, Oval represents end of the process and arrows show the flow of controls. All shapes represent UML notations.

```
Procedure Mining Process ( )
{
    1- Define the problem statement
    2- While Accuracy Acc is less than a predefined Threshold T.
        a. Gather the Requirements of the Solution
        b. If (Data is available)
            i. Create Database DB.
            ii. If(AlgoS)
                1. If(AlgoS ≠ AlgoR)
                    a. R = Requirements Gathering(AlgoS)
                    b. AlgoU = Modify(AlgoS , R)
                    c. Acc = Implement(AlgoU)
                    d. Update toping condition of Acc(Acc<T)
                2. Else
                    a. Acc = Implement(AlgoS)
                    b. Update toping condition of Acc(Acc<T)
            iii. Else
                1. Create New Algorithm AlgoN
                2. Acc = Implement(AlgoN)
                3. Update toping condition of Acc(Acc<T)
        c. Else
            i. Go to step. a
    3- End While
    4- Return
}
```

Fig 3.2. Procedure of framework workflow.

3.1 Defining Problem Statement

Defining problem statement Process is the most important step, which comes in any research development. It is very important to clearly define the problem statement before moving toward next step. We are defining the problem statement in the area of gas production.

Pakistan is blessed with many energy resources such as natural gas, coal, thermal, hydro, solar, and oil energy but due to less technological development and the poor management in utilizing these resources in the cost-effective way. The shortage of energy crises started in the country rapidly since last ten years. Further, rapid increase in the number of customers and change in the life style of consumers caused a rapid ripple effect in the pattern of consumption of energy requirements. This effect became a big issue in the country's economy. In this situation, it is needed to come up for a future plan to work out the customer's energy demand and energy production/ availability. To solve this energy problem the authors decided to come up with an efficient forecasting model and handle the problem step by step. We have decided to handle the natural gas energy problem first because natural gas energy is the cheapest, easily available and most demanding energy source in the country, and Pakistan has a very good natural gas production technology and supply infrastructure. In

this research authors developed a prediction model by using the novel techniques of Data Mining and Artificial Intelligence in which availability of large amount of historical data is pre-requisite.

This approach will be helpful to the decision makers in strategic decision-making. We successfully able to get the historical data of twenty one years starting from 1989 to year 2010 coming from 26 different gas production fields and this data will be useful to apply in the predication model. In our project we are working to answer the question “In the past 21 years what was the gas production trends in the country” and “how to control the gas production to avoid the gas energy crises?” Further, we will determine projections of the gas energy production so that resources may be utilized optimally to achieve the maximum advantage and evaluate which technique suits best to the available type of data?

3.2 Solution Requirements Gathering

In the proposed framework the next step is “solution requirements process”. Authors proposed preference learning agents within this process to make it more efficient and effective. Preference learning agents learn the preferences of the user through the output of the problem definition process and make the solution requirements gathering process efficient.

To understand the actual requirements and needs we did extensive efforts for data collection tasks to find actual facts and analyze the current situation of the energy crises in the country. In this paper authors are presenting the following requirements which our gas prediction model will fulfill by extracting information of the gas energy production and consumption at various locations, to analyze the behavior and also predict the future trends. We can summarize the solution requirement of this research in following points

1. Future production trends of Natural gas energy.
2. Future consumption trends of Natural gas energy.
3. Cost effective decision making facts for the gross domestic product (GDP).
4. Data presentation and analyzing views of historical and current trend of Natural gas energy production.

3.3 Data Availability Checking Process

The quality, dimensionality and detail of data are very important to analyze and mine novel information hidden in the data. In this step the AI agent; data gathering agent’ is involved by the authors, which helps in collecting the relevant data efficiently.

To mine the historical data on natural gas production authors visited different famous and big companies in the country e.g. Federal Bureau of Statistics (FBS), Hydrocarbon Development Institute of Pakistan (HDIP) and Oil & Gas Development Company Limited (OGDCL) and many more. By having a preliminary survey the authors decided to consider the data of OGDCL. The reason for considering the data was the availability of the detailed structured data. The OGDCL provided the monthly data of last 21 years (1989-2010), which was about 26 locations of the country.

The fig.3.1 shows there is a possibility of finding the appropriate data or not. Data mining process’s the historical data for analyzing the future trends. If the research found the required data then they can move forward towards step 3 in the framework. If the required data is not available, due to any reason, then researcher should take a step back and start again from the step 2 of “solution requirement’s process”. As if the data is not available it’s not possible to achieve exactly the required solution therefore the solution requirements defined in step2 needs to be redefined. After refining the solution requirements perform the step 3 of “data availability checking process” again.

3.4 Database Creation Process

Database Creation Process is a vital task because the database is itself the most important part of any forecast model. Its creation is critical because the future production model will work by using the data from the targeted database. Therefore, it is very important to work with good planning and full concentration on designing of the database. During the design of database the first step is the selection of the type of database which must be according to the nature of the data and required data

which the application will use. A simple relational database with ER model, a data warehouse with dimensional modeling or a data mart with dimensional modeling can be used in this type of the projects. During the process of finalizing the type of database the nature of the available data has to be considered carefully because it is a radical factor of influence in making the forecasting modal. In our project the nature of the data used is detailed, historical and growing in nature. The authors consider the data mart as a suitable database engine; the data mart is entitled as “Energy Analytical Data Mart”.

The second step in database creation process is defining the dimensions of the data to be considered. The dimension should directly or indirectly relate to the targeted fact of the prediction model. The Dimensionality Modeling is a logical design technique to presents the data in a standard, intuitive form for high-performance access. Every dimensional model (DM) is composed of one table with a composite primary key, called the fact table, and a set of smaller tables called dimension tables. Each dimension table has a simple, non-composite primary key that corresponds exactly to one of the components of the composite key in the fact table [4]. Particularly Star Schema is used for dimensionality modeling of the data to design the “Energy Analytical Data Mart” due to the nature of the gathered data i.e. not hierarchal distributed. Star Schema is the best approach which we have considered. In our project the data is divided into four dimensions, namely Product, Time, Employee and Area plus one fact table which contains the foreign keys of all four dimensional tables and historical facts. The gathered data is converted into required data structure for the prediction model. Fig.3.3 below shows the star schema of the implemented “Energy Analytical Data Mart”.

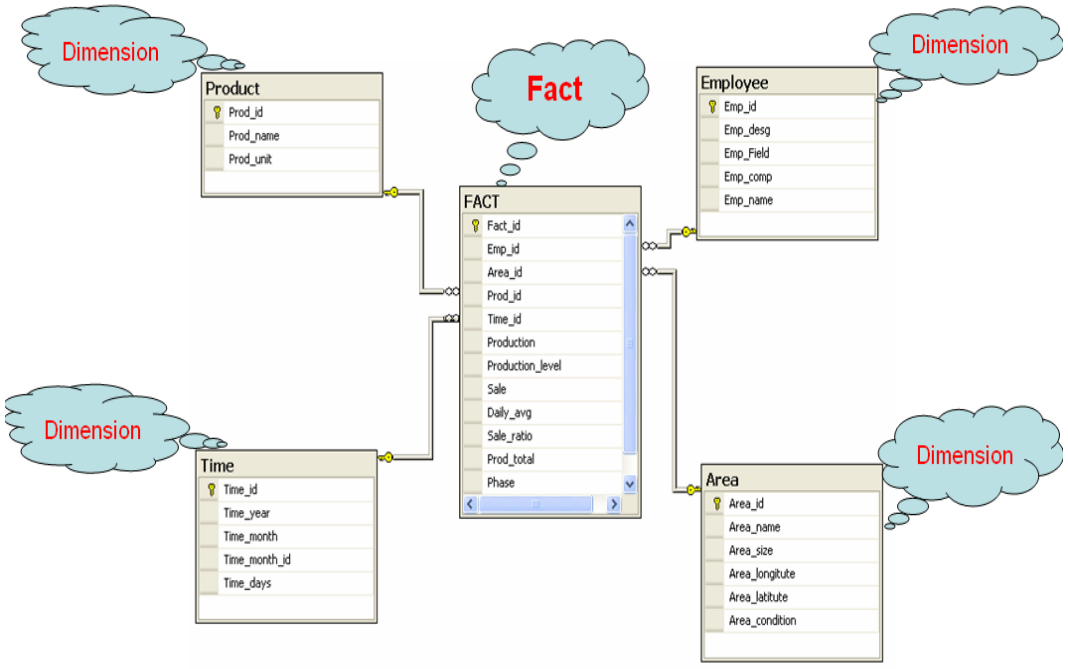

Fig 3.3. Star Schema of the Energy Analytical Data Mart for prediction model of Natural Gas Energy Production.

3.5 Algorithm Selection Process

After the completion of database creation and defining the parameters in step.4, the next step we have followed is the step.5 of selection of suitable algorithm for this type of problem. For solving the gas production forecasting problem, authors decided to select two algorithms after doing a comparative analysis of all existing algorithms in literature. Later the algorithms would be implemented and evaluated on the basis of data that is to be mined, the required outputs and processing efficiency.

After analyzing the efficiency of neural networks with evolutionary algorithms, authors selected the artificial neural networks with evolutionary algorithms. In this research the authors consider Artificial

Neural Networks (ANN) to model gas production in the country because of its efficiency observed in literature.

During the step.5 of algorithm selection process there is the possibility of not finding the perfectly fitted algorithm as per requirement. If the researchers are unable to find any appropriate algorithm then they can move on to the step.6 of "Designing a new algorithm process" and invent a new solution as per the requirements defined in step.2.

3.6 New Algorithm Creation Process

This step will be followed in case of failure in finding an appropriate existing algorithm during the step.5 of "Algorithm Selection Process". In this step researcher have to create a new algorithm which fulfills the requirement of the solution they want with better efficiency in producing accurate output. This step.6 of "New Algorithm Creation Process" is time consuming and needs expert's knowledge. To make it more easy and efficient authors placed an AI agent for modeling. Modeling agent helps in modeling the problem's solution into algorithm efficiently.

3.7 Changes Requirement Check

In case of finding an appropriate mathematical solution in step.5 "Algorithm Selection Process" of as per requirements defined in step.2, the researchers should analyze that the found algorithm needs any changes in the algorithm or not. If changes are not required researcher can move to step.10 of "Mining Process" directly and can skip the steps 8 and 9 of the "Changes Requirements Process" and the "Algorithm Updating Process" respectively. In most of the cases all the existing algorithm needs some changes as per the requirements of research domain. In case of changes required in the selected mathematical model, the researcher should follow the step.8 of "Changes Requirements Gathering Process".

3.8 Changes Requirements Gathering Process

In the step.7 of "Changes Requirements Gathering Process" the researchers have to finalize the steps and functions, which need to be changed in the existing algorithm. Researchers have to finalize the requirements and list up them with mutual decisions of experts and also by the help of artificial agent which learn the preference of the user for modified algorithm. All the work should be done with proper documentation to facilitate from that documentation in the next step.9 of "Algorithm Updating Process".

3.9 Algorithm Updating Process

After finalizing the changes requirements in step.8, in this step researchers should write down the full updated algorithm/mathematical model in appropriate structural format and review the whole process of algorithm as a dry run. The main objective of this step is to apply changes in the algorithm as per requirements gathered in last step and recheck the whole algorithm process before going to implementation.

3.10 Implementation Process

Step.10 of "Implementation process" is the most technical and critical step after the step.4 of "Database creation". In this step researchers need the appropriate development frameworks and full command on the implementation process of mathematical model or Algorithm formed in last step.9. This is the step of transforming mathematical model into implemented predictive model. In this process modeling agents can help and guide the whole process.

3.11 Accuracy Achievement Check

After the creation of prediction model it is very important to check the functionality of the model. Researchers have to do brief analysis by doing several experiments. On the basis of this analysis researcher or group of researchers will decide that this is the model they were looking forward to achieve after this research exercise or not. If the aimed results are not achieved then researcher can restart their research exercise from step.5, step.3 or even from step.2. From where to restart the research exercise, depends on the percentage of accuracy they achieved. If researchers achieved the

goals they were aiming then this is the prediction model and final outcome of their research, which is achieved with efficiency.

4 CONCLUSION AND FUTURE WORK

Data mining is a powerful tool for the analyses of historical data and mine the data to achieve the required target. Data mining provide better opportunity to take right decisions at the right time in real time scenarios. But due to unavailability of a complete and well defined framework to mine the target, the researchers find difficulties. In this research paper authors presents an efficient conceptual framework with its complete flow of processes. All the processes are described with their details. With the help of algorithm authors have presented sequence of all the steps. This research work provides an efficient path to any new research in the field of data mining especially in the predictive data mining.

In this research work it can be observed that the authors have discussed the natural Gas Energy problem in Pakistan only till the step of “Algorithm Selection Process”. The reason of not discussing the problem for further steps is because the further research is under process. Authors are doing research on the Natural gas energy forecasting problem statement in light of the presented framework in this paper. In future full analyses of this framework using Natural gas energy forecasting problem will be presented catering the limitations related to depletion of known sources of Natural Gas and probable addition of its new resources including Import of Natural Gas from other countries. The future outlook model thus proposed in the forth coming work will be validated by using some feasible approach.

5 REFERENCES

[1] Alam Nayyar, “Renewable hot dry rock geothermal energy source and its potential in Pakistan”, Renewable & Sustainable Energy Reviews, Elsevier, 2010, Pp. 1124-1129

[2] Tahir, “Power Crisis in Pakistan - What’s the solution?”, available at: http://empowerpakistan.blogspot.com/2008/05/energy-crisis-and-pakistan.html, Reviewed on MAY 28, 2008.

[3] Munawar A. Sheikh, “Energy and renewable energy scenario of Pakistan”, Renewable & Sustainable Energy Reviews, Elsevier, 2010, Pp 354-363.

[4] Thomas Connolly and Carolyn Begg, P.E.L, DATABASE SYSTEMS, 5th Ed, Pearson Education Limited, Edinburgh Gate Harlow, Essex CM20 2JE, England, (2009).

[5] P.S.Bradley, Usama M. Fayyad, O.L. Mangasarian. ”Data Mining: Overview and Optimization Opportunities”. Journal of Computing, special issue on Data Mining. January 19, 1998.

[6] Mehta, R. Agrawal, and J. Rissanen. Sliq: a fast scalable classifier for data mining. In Proceedings of EDBT-96. Springer Verlag, 1996.

[7] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and I.C. Verkamo. Fast discovery of association rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in knowledge Discovery and Data Mining, pages 307 { 328. MIT Press, Cambridge, MA, 1996.

[8] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: A new data clustering algorithm and its applications. Data Mining and Knowledge Discovery, 1997.

[9] H. Mannila, H. Toivonen, and A.I. Verkamo. “Discovery of frequent episodes in event sequence”, Data Mining and Knowledge Discovery, 1997.

[10] G. Piatetsky-Shapiro and W. Frawley, editors. Knowledge Discovery in Databases. MIT Press, Cambridge, MA, 1991.

[11] Glymour, R. Scheines, and P. Spirtes ABD K. Kelly. Discovering Causal Structure. Academic Press, New York, 1987.

[12] D. Heckerman. Bayesian networks for data mining. Data Mining and Knowledge Discovery, 1997.

[13] R.O. Duda and P.E. Hart. Pattern Classification and Scene Analysis. John Wiley and Sons, New York, 1973.

[14] Munir H. Naveed, Sheikh S. Ahmad, Sobia Khalid and Sehresh Khan, "Development of Prediction Model for the Concentration Level of Air Toxin in the city of Rawalpindi Using Artificial Neural Network", World Applied Sciences Journal, Aug 2010.

[15] Fangwen Zhai, Qinghua Wen, Zehong Yang, Yixu Song, “Hybrid Forecasting Model Research on Stock Data Mining” , 4th International Conference IEEE, May 2010.

[16] Usama Fayyad, Gregory Piatetsky-shapiro and Padhraiz smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework”, AAAI, 1996.

[17] Rie Kubota Ando and Tong Zhang, “A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data”, J. Mach. Learn. Res, Dec 2005, Pp. 1817-1853.

[18] Mitra, S., Pal, S. K. and Mitra, P., “Data mining in soft computing framework: a survey IEEE Transactions on Neural Networks”, 2002, pp. 3-14. ISSN 1045-9227.