

## THREE PHASE ITERATIVE MODEL OF KDD

Simmi Bagga<sup>1</sup> & G. N. Singh<sup>2</sup>

---

KDD (Knowledge Discovery in Data Bases) is multi-step process. KDD is the process that navigation to knowledge from the data base according to user requirements. This paper mainly explains the concept of three phase iterative model of KDD. It elaborates the traditional process of KDD and also explains the various limitations in that process. This paper also explained the process of three phase iterative model of KDD that can overcome some of the limitations caused in traditional process.

Keywords: 8KDD, Traditional Model, Iterative Model

---

### 1. INTRODUCTION

KDD process includes how data is stored and accessed, how and what algorithms can apply to large amount of data efficiently, how results are interpreted and visualized. KDD is the process of identifying valid, useful and understandable patterns in data. Knowledge discovery is defined as the non trivial process of identifying valid, useful, understandable and interesting information from the large data stored in the database.

### 2. TRADITIONAL KDD PROCESS

Knowledge is a pattern that is sufficiently interesting to the user. The user specifies the measure of interestingness and the certainty criterion. Discovered knowledge is the output from a program that analyzes a data set and generates patterns. A pattern's certainty is the measure of confidence in discovered knowledge represented by the pattern. The Traditional KDD Process is:

#### A. Task Analysis

The first step of KDD is defining the application domain and the goal of the whole process. In this step we mainly specify the problem specifications. On the basis of requirement we recognize the patterns of data then we formulate strategies for determining the knowledge that fulfil the requirement.

#### B. Data Selection

After specifying the problem specification we select and prepare a data set for process from the large database that data is usually refers to as a target data. We select the target data because KDD process is never done over the entire

database. Selection is an appropriate procedure for generating the target data set from the database.

#### C. Data Cleaning

Data cleaning is removing bad data, but also finding hidden correlations in the data, identifying sources of data that are the most accurate, and determining which columns are the most appropriate for use in analysis. Dirt may also include missing data information, duplicate data etc. Data cleaning process involves the examining of data for completeness and integrity and it's a very time consuming and labour intensive. Pattern matching algorithm helps in data cleaning.

#### D. Data Transformation

After cleansing the data is transformed in the different format that is known as data transformation. The data is converted in to a form that is effectively used in the data mining step.

#### E. Data Mining

Data Mining is one of the main steps of the whole KDD process. Data mining is used to extracting patterns from data. Data mining required cleaned, transformed in proper format and coded data. Data mining is the core process that takes input cleaned & transformed data and searches patterns using some algorithms and then results patterns and relationships. There are various types of data mining algorithms that can be used during the process are: classification, regression, clustering, etc.

#### F. Pattern Interpretation & Evaluation

In this step results has to be interpreted and evaluated to discover knowledge from the patterns. The pattern extraction collects the algorithms to extract the patterns. The evaluation evaluates the interests and utilities of extracted patterns. In this step we also visualize of extracted patterns and models and also visualize of the data given the extracted models.

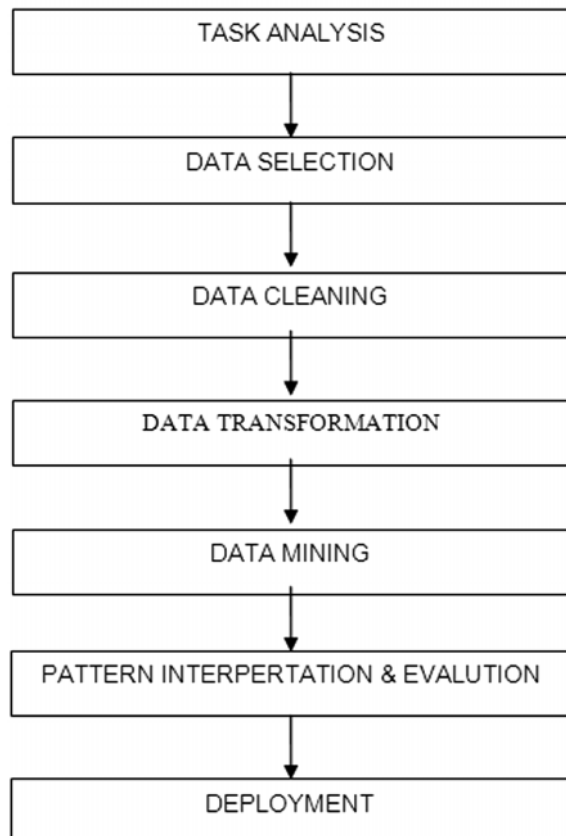
---

<sup>1</sup>Department of Computer Science, Sant Hira Dass Kanya MahaVidyalaya, Kala Sanghian, Distt Kapurthala, Punjab (India)

<sup>2</sup>Department of Physics and Computer Science, Sudarshan Degree College, Lalgaon Distt. Rewa (M.P.)(India)  
E-mail: [1simmibagga@gmail.com](mailto:1simmibagga@gmail.com)

## G. Deployment

In this step of KDD the final report & presentations are prepared. Update the models after review and analysis if required. Deployment-creation of the model is generally not the end of the project.



### 3. LIMITATIONS IN THE TRADITIONAL KDD PROCESS

Data Mining is the main step of KDD. This step is only successful when task discovery & data selection steps are valid i.e. if it meets the requirement of the user. After deciding the target data we have to validate whether the quality and quantity of data is satisfactory for the goal. It is very important that the target data must be verified otherwise the whole process may fail. In short we can say it deals to find the prior knowledge and the setting the goal of the application process. Data mining operation may generate thousands of patterns but all of them are not interesting. The main Limitations in the traditional KDD Process are:

- The biggest disadvantage of the Traditional process is that we cannot go back, if the any of the previous step has gone wrong, things can get very complicated & the whole process may get fail.
- Any error in data selection or pattern selection can cause a lot of problem and if we recognize this problem in the later stages we can overcome it.

- High amounts of risk and uncertainty in resulting knowledge because from the initial stages we can't relay whether the resulting knowledge is according to the requirement or not.
- Handling errors in the previous stages are expensive to fix and new requirements are expensive to incorporate in later stages.

### 4. CONCEPT OF THREE PHASE ITERATIVE MODEL OF KDD

KDD is multi-step process. The KDD process is interactive and iterative, involving various steps with many decisions being made by the user. KDD refers to the overall process of discovering useful knowledge from data. In the traditional KDD Steps if one of the phase take wrong decision then whole process will fail. If we modify this process by providing reverse feedback then the resulting knowledge will be more accurate than the traditional approach. In three phases iterative model of KDD the whole process is divided into three stages that are: Pre-processing, Data Mining and Post Processing.

#### A. Pre-processing

This step mainly involve data selection according to the requirements, cleaning of Data that involves the examining of data for completeness and integrity then data is transformed in the different format that is known as data transformation. This is one of the main step of the Data mining because the success of whole process depends upon this step. While performing this step, if any case we identified that the selected data set can not provide the accurate result then we can change this data set because this step can iterate. But in the case of traditional model if the wrong data set has been selected then the whole process can fail and the whole process has to start again in that case. The pre-processing step three phase iterative model repeats till we ensure the selected pattern will give the required result. The output of this step is only to select the subset of data that contain the information for which we are performing whole process. Once we get the appropriate subset we can move to the next step of the model i.e. Data Mining.

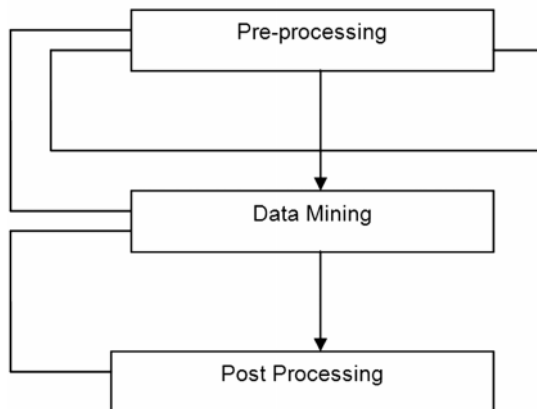
#### B. Data Mining

Data mining is the iterative and interactive involves the various steps with the decision made by the user. Data mining is only the one stage in KDD process concerned with applying computational techniques to find patterns discovery in a data set from which noise has been previously eliminated and which has been transformed in such a way to enable the pattern discovery process. This step will get only successful if the previous step performed accurate. If the any error occur in the initial stage the result of whole KDD process will not be feasible.

### C. Post Processing

After completing Data Mining, we visualized the results. Visualization plays an important role in making discovered knowledge to understand and interpret by humans.

It's the last phase of the KDD process. In this step results has to be interpreted and evaluated to discover knowledge from the patterns. Even the purpose of the model is to increase knowledge of the data, the knowledge will need to be organized and presented in a way that the customer can use it.



The goal of both the models is same i.e. to extract knowledge to meet the requirement of the user but the working of both the model is totally different. In the traditional model whole process of KDD is divided in to seven steps and there is not any backward move to the previous step and if any problem detect in the later steps then there is no chance to recover that error. In the three phase iterative model of KDD process the whole process is divided in to three steps i.e. pre-processing, Data mining and Post Processing. The pre-processing step involves task discovery, data selection, data cleaning and data transformation. The result of this stage is pattern selection to which we perform Data mining step. This model also provides the backward link from data mining step to pre-processing. If any error or changes are required in the pattern

selection while performing mining process we can select again by moving backward.

### 5. CONCLUSION

KDD is used to extract knowledge from data. Two different models have been explained in this paper. One is the traditional model, in which whole process is divided into seven steps and other one is three phase iterative model of KDD in which the process is divided into three steps and each step has forward as well as backward link with each other. Three phase iterative model of KDD provides us more accurate result than the traditional model.

### REFERENCES

- [1] Fayyad, U.M, "Data Mining and Knowledge Discovery: Making Sense Out of Data", IEEE Expert 11(5), 1996.
- [2] T. Imielinski and H. Mannila, "A Database Perspective on Knowledge Discovery", Communications of ACM, 39, 1996.
- [3] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.
- [4] Elder, John F., IV and Daryl Pregibon, (1996), "Advances in Knowledge Discovery & Data Mining", A Statistical Perspective on KDD".
- [5] Chen, M. et al, 1996, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, 8.
- [6] Han, J., & Kamber, M. (2006), "Data Mining Concepts and Techniques", Boston: Elsevier.
- [7] Dunham, M.H. (2003), "Data Mining Introductory and Advanced Topics", Upper Saddle River, NJ: Pearson Education, Inc.
- [8] Yao, Y.Y., "A Step Towards the Foundations of Data Mining, In: Data Mining and Knowledge Discovery: Theory, Tools, and Technology V, Dasarathy, B.V. (ed.)", The International Society for Optical Engineering.