International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 12 No. 1 Nov. 2014, pp. 217-222

http://www.ijisr.issr-journals.org/

A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)

Umair Shafique and Haseeb Qaiser

Department of Information Technology,
University of Gujrat,
Gujrat, Pakistan

ABSTRACT: Data Mining is about analyzing the huge amount data and extracting of information from it for different purposes. From the last few years the field of Data Mining becomes prominent and makes huge growth. There are different standard models for data mining. All these models are defined in sequential steps. These steps help in implementing the data mining tasks. In this paper we will compare these models and give brief understanding about them.

KEYWORDS: Knowledge, Applications, Patterns, Modeling, Information.

1 INTRODUCTION

Over the last few years the field of data mining [1] becomes very important for different industries, co-corporations and businesses etc because of its ability to use huge amount of data that had previously no use and makes analysis and predicting trend and patterns. Basically the risk of wasting the wealthy and valuable information contained by the big databases was arise and this requires the use of adequate techniques to get useful knowledge [2] so that the field of data mining had been emerged in 1980's and is still making progress. With the emergence of this field different process models were introduced. These process models guides and carry the data mining tasks and its applications.

Efforts were made to use data mining process models that may guide the implementation of data mining on big or huge amount of data. In our paper we mainly focuses on three most popular data mining process models and these models are Knowledge Discovery Databases (KDD) process model, CRISP-DM and SEMMA. These three models are mostly practiced by the data mining experts and researchers. In this paper we will elaborate these models. We will do comparative study of these models and try to explain that what insight of them is.

2 OVERVIEW OF KDD, CRISP-DM AND SEMMA

The Knowledge Discovery Databases (KDD) model is an iterative and interactive model [3]. It has total nine steps. It refers to finding knowledge in data and emphasizes the high level of specific data mining method.

Cross-Industry Standard Process for Data Mining (CRISP-DM) [4] was launched in late 1996 by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR. This models the refines over the years. It has six steps or phases.

Sample, Explore, Modify, Model, Assess (SEMMA) [5] model was developed by SAS institute. It has five different phases.

2.1 THE KDD PROCESS MODEL

The KDD or Knowledge Discovery Databases [6] is the process of extracting the hidden knowledge according from databases. KDD requires relevant prior knowledge and brief understanding of application domain and goals. KDD process model is iterative and interactive in nature. There are nine different steps or stages of this model and these are given below.

Corresponding Author: Umair Shafique

2.1.1 DEVELOPING AND UNDERSTANDING OF THE APPLICATION DOMAIN

This is the first stage of KDD process in which goals are defined from customer's view point and used to develop and understanding about application domain and its prior knowledge.

2.1.2 CREATING A TARGET DATA SET

This is the second stage of KDD process which focuses creating on target data set and subset of data samples or variables. It is an important stage because knowledge discovery is performed on all these.

2.1.3 DATA CLEANING AND PRE-PROCESSING

This is the third stage of KDD process which focuses on target data cleaning and pre-processing to complete and consistent data without any noise and inconsistencies. In this stage strategies are develop for handling such type of noisy and inconsistent data.

2.1.4 DATA TRANSFORMATION

This is the fourth stage of KDD process which focuses on transformation of data from one form to another so that data mining algorithms can be implemented easily. For this purpose different data reduction and transformation methods are implemented on target data.

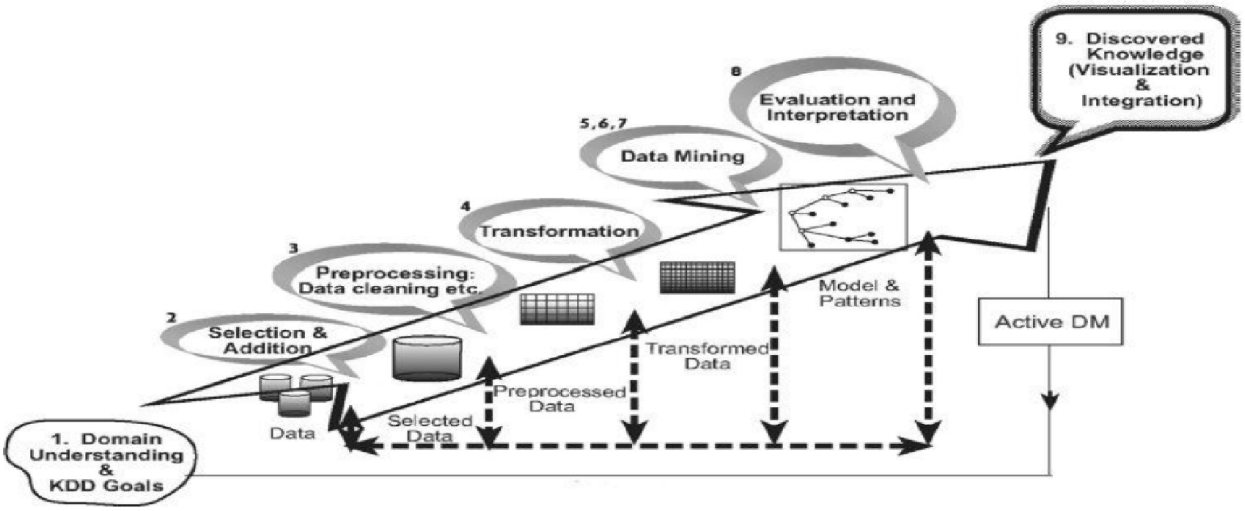

Fig. 1. Knowledge Discovery Databases (KDD) Process Model

2.1.5 CHOOSING THE SUITABLE DATA MINING TASK

This is the fifth stage of KDD process in which appropriate data mining task is chosen based on particular goals that are defined in first stage. The examples of data mining method or tasks are classification, clustering, regression and summarization etc.

2.1.6 CHOOSING THE SUITABLE DATA MINING ALGORITHM

This is the sixth step of KDD process in which in which one or more appropriate data mining algorithms are selected for searching different patterns from data. There are number of algorithms present today for data mining but appropriate algorithms are selected based on matching the overall criteria for data mining.

2.1.7 EMPLOYING DATA MINING ALGORITHM

This is the seventh step of KDD process in which selected algorithms are implemented.

2.1.8 INTERPRETING MINED PATTERNS

This is the eighth step of KDD process that focuses on interpretation and evaluate of mining patterns. This step may involve in extracted patterns visualization.

2.1.9 USING DISCOVERED KNOWLEDGE

This is the last and final step of KDD process in which the discovered knowledge is used for different purposes. The discovered knowledge can also be used interested parties or can be integrate with another system for further action.

2.2 THE CRISP-DM PROCESS MODEL

CRoss-Industry Standard Process for Data Mining (CRISP-DM) was developed by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR in 1999, CRISP-DM 1.0 version was published and is complete and documented. It provides a uniform framework and guidelines for data miners. It consists of six phases or stages which are well structured and defined [7]. These phases are described below.

2.2.1 BUSINESS UNDERSTANDING

This is the first phase of CRISP-DM process which focuses on and uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.

2.2.2 DATA UNDERSTANDING

This is the second phase of CRISP-DM process which focuses on data collection, checking quality and exploring of data to get insight of data to form hypotheses for hidden information.

2.2.3 DATA PREPARATION

This is the third phase of CRISP-DM process which focuses on selection and preparation of final data set. This phase may include many tasks records, table and attributes selection as well as cleaning and transformation of data.

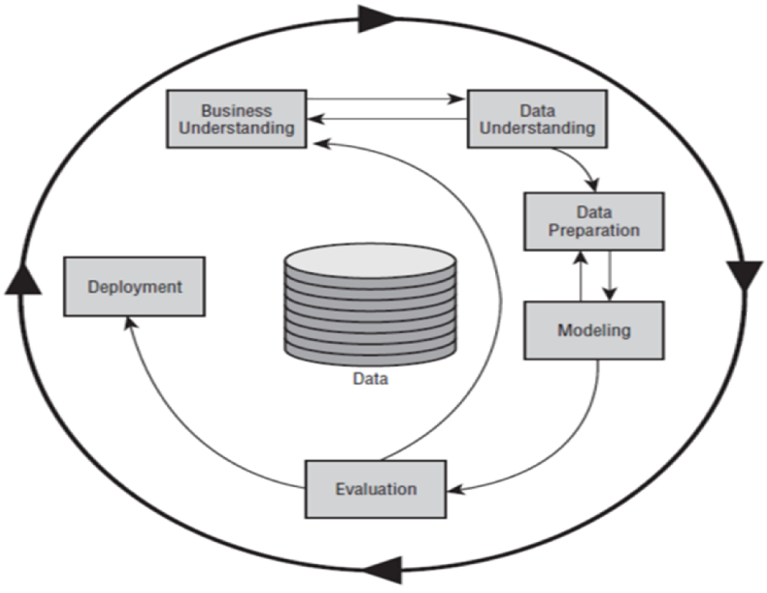

Fig. 2. CRISP-DM Process Model

2.2.4 MODELING

This is the fourth phase of CRISP-DM process selection and application of various modeling techniques. Different parameters are set and different models are built for same data mining problem.

2.2.5 EVALUATION

This is the fifth stage of CRISP-DM process which focuses on evaluation of obtained models and deciding of how to use the results. Interpretation of the model depends upon the algorithm and models can be evaluated to review whether achieves the objectives properly or not.

2.2.6 DEPLOYMENT

This is the sixth and final phase of CRISP-DM process focuses on determining the use of obtain knowledge and results. This phase also focuses on organizing, reporting and presenting the gained knowledge when needed.

2.3 THE SEMMA PROCESS MODEL

The SEMMA stand for (Sample, Explore, Modify, Model, and Access) is data mining method developed by SAS institute. It offers and allows understanding, organization, development and maintenance of data mining projects. It helps in providing the solutions for business problems and goals. SEMMA is linked to SAS enterprise miner and basically a logical organization of the functional tools for them. It has a cycle of five stages or steps.

2.3.1 SAMPLE

This is the first and optional stage of SEMMA process which focuses on sampling of data. A portion from a large data set is taken that big enough to extract significant information and small enough to manipulate quickly.

2.3.2 EXPLORE

This is the second stage of SEMMA process which focuses on exploration of data. This can helps in gaining the understanding and ideas as well as refining the discovery process by searching for trends and anomalies.

2.3.3 MODIFY

This is the third stage of SEMMA process which focuses on modification of data by creating, selecting and transformation of variables to focus model selection process. This stage may also looks for outliers and reducing the number of variables.

2.3.4 MODEL

This is the fourth stage of SEMMA process which focuses on modeling of data. The software for this automatically searches for combination of data. There are different modeling techniques are present and each type of model has its own strength and is appropriate for specific situation on the data for data mining.

2.3.5 ACCESS

This is the fifth and final stage for SEMMA process focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.

3 COMPARATIVE ANALYSIS OF KDD, CRISP-DM AND SEMMA

There are nine, six and five stages for KDD, CRISP-DM and SEMMA process model respectively. By examining all the three data mining process models they clearly shows that they are somehow equivalent to each other even SEMMA is directly linked to the SAS enterpriser miner software and CRISP-DM Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR. The comparison between them shows that

- The KDD process step "Developing and Understanding of the Application Domain" can be identified with "Business Understanding" phase of CRISP-DM process.
- The KDD process steps "Creating a Target Data Set" and "Data Cleaning and Pre-processing" can be identified with "Sample" and "Explore" stages of SEMMA respectively, and/or these can be identified with "Data Understanding" phase of CRISP-DM.
- The KDD process stage "Data Transformation" can be identified with "Data Preparation" stage of CRISP-DM and "Modify" stage of SEMMA process respectively.
- The three stages of KDD "Choosing the suitable Data Mining Task", "Choosing the suitable Data Mining Algorithm" and/or "Employing Data Mining Algorithm" can be identified with "Modeling" phase of CRISP-DM and/or "Model" stage of SEMMA process respectively.
- The KDD process step "Interpreting Mined Patterns" can be identified with "Evaluation" phase of CRISP-DM process and/or "Assessment" stage of SEMMA process respectively.
- The KDD step "Using Discovered Knowledge" can be identified with "Deployment" phase of CRISP-DM process.

The comparison of these models is given in below table.

Table 1: Summary of KDD, CRISP-DM and SEMMA Processes

Data Mining Process Models	KDD	CRISP-DM	SEMMA
No. of Steps	9	6	5
Name of Steps	Developing and Understanding of the Application	Business Understanding	----------
	Creating a Target Data Set	Data Understanding	Sample
	Data Cleaning and Pre-processing		Explore
	Data Transformation	Data Preparation	Modify
	Choosing the suitable Data Mining Task	Modeling	Model
	Choosing the suitable Data Mining Algorithm		
	Employing Data Mining Algorithm		
	Interpreting Mined Patterns	Evaluation	Assessment
	Using Discovered Knowledge	Deployment	----------

4 CONCLUSION

Our study was on comparison between KDD, CRISP-DM and SEMMA data mining processes. As we know most of the researchers and data mining experts follow the KDD process model because it is more complete and accurate. In contrast CRISP-DM and SEMMA or mostly company oriented especially SEMMA that is used by SAS enterprise miner and integrate with their software. However, study shows that CRISP-DM is more complete as compare to SEMMA. All and all these process models guides and helps the people and experts to know that how they can apply data mining into practical scenarios.

ACKNOWLEDGMENT

We will like to thank our Parents and Family members. It is due to their support and guidance that encourage us to write this paper.

REFERENCES

[1] Han, J. and Kamber, M. "Data Mining: Concepts and Techniques. Second Edition", Morgan Kaufmann Publishers, San Francisco, 2006.

[2] Chen, M. et al, "Data Mining: An Overview from a Database Perspective.", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp 866-883, 1996.

[3] Brachman, R. J. & Anand, T., "The process of knowledge discovery in databases.", AAAI Press / The MIT Press. 1996.

[4] Chapman, P. et al, "CRISP-DM 1.0 - Step-by-step data mining guide." SPSS, 2000.

[5] SAS Enterprise Miner – SEMMA. SAS Institute, 2014 [online] available: http://www.sas.com/technologies/analytics/datamining/miner/semma.html (September 2014.)

[6] Usama Fayyad et al., "From Data Mining to Knowledge Discovery in Databases" American Association for Artificial Intelligence, 1996.

[7] Colin Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining", JOURNAL of Data Warehousing, Volume 5, Number 4, page. 13-22, 2000.

AUTHOR'S BIOGRAPHY

UMAIR SHAFIQUE- Pursuing M.Sc degree in Information Technology, Session 2012-2014, from Department of Information Technology, University of Gujrat, Gujrat Pakistan.

HASEEB QAISER- Pursuing M.Sc degree in Information Technology, Session 2012-2014, from Department of Information Technology, University of Gujrat, Gujrat Pakistan.