ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

# The Prophecy-Prototype of Prediction modeling tool

Ms. Ashwini Dalvi[1], Ms. Dhvni K.Shah[2], Ms. Rujul B.Desai[3], Ms. Shraddha M.Vora[4], Mr. Vaibhav G.Tailor[5]

Department of Information Technology, Mumbai University, Mumbai, India

**Abstract--Now a days, a lot of data mining tools and software have been developed. Due to the availability of so many tools, there is a need for the user to select the tool appropriate for his need. This paper provides with the tool selection criteria, the categories of data mining tools, summarization of category and the overview of existing tools in various fields. This paper also comments on usability limitation of existing data mining tools. In this paper we proposed prototype of prediction tool "The Prophecy". This tool focuses only on prediction of data and assumes that the input data is already clean. The purpose of this tool is to enable any type of user, with technical knowledge of data mining or without knowledge of data mining to see prediction of future based on his data.**

**Keywords--Data mining tools, Prediction modeling, Usability Limitation, The Prophecy**

## I. INTRODUCTION

Today, in the current era, for every big or small company/industry/business it is very important for them to know each and every data they have with them. Knowing the data means, What is the data? What kind of data is present in the company's database? How is the data collected? What is the nature of data? For all this questions the answer is Data Mining. The data present in the company's database may be important for making strategic plans for the growth of the company. However, this data becomes useful only when cleaned and analysed because raw data cannot be directly used to come to any conclusion. The field of analysis of this data is known as data mining.

Data mining is an integral part of Knowledge Discovery in Database (KDD) (Refer figure 1). Sometimes, data mining is used interchangeably with KDD.

The data mining process includes Defining the problem, Preparing data, Exploring data, Building models, validating models, Deploying and updating models.

Data mining tools are very popular in medical, education, banking, real estate and retail industry because there is a large amount of data present which can be put to use for analysis. With the growing amount of data its analysis becomes difficult but also very useful and hence data mining is so important in the above mentioned domains. In medical domain the patient's historical information about the symptoms help in identifying the diseases in future patients using data mining.. In the retail industry with ever increasing competition, forecasting the future sales and managing inventory helps in boosting the sales. In education sector learning data mining is a subject of interest for the students.

Due to rapid advancements in each of the respective field, there are a variety of tools available and hence the selection of the most appropriate tool is of utmost importance. Selecting a wrong tool can cost the user in terms of money, time as well as other important data.

The paper is further divided in the following sections:
II. Tool Selection Criteria
III. IV. Usability limitations of existing tools
IV. Proposed prototype of prediction tool
V. Conclusion

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

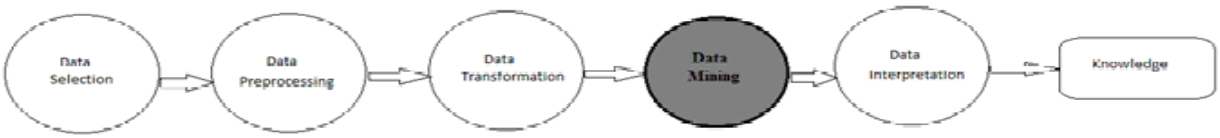


Figure 1. Knowledge Discovery in Database Process

## II. TOOL SELECTION CRITERIA

The research has been conducted to comment on appropriate data mining technique for various domains such as Health Care, Education, Retail, Banking etc. Some approaches had been suggested to evaluate the data mining techniques for different data mining tasks such as Classification, Prediction [1]
The following section gives the categorization of Data Mining tools to different types [2][3][4]:

- *Traditional Data Mining Tools:*

They use Complex Algorithms. Some tools are installed on Desktop. Some capture information from outside the Database. These tools are supported by Windows and Unix. Majority of tools specialize in one Operating System only. They handle data using OLAP.

- *Dashboards:*

They reflect data changes and updates on the screen. They are installed on computer and they show the changes and updates in the database in the form of charts, tables (displayed on the screen). It enables the users to see the Business performance. It also compares historical and current data to see the changes happening in business performance.

- *Text Mining Tools:*

Initially text mining was done only on structured data but now current data mining techniques can perform mining on semi-structured and unstructured data as well. These tools are used to combine various text types like Microsoft Word, Acrobat PDF and other text files into a single type and then analysis can be done on this text processing. The database support can range from weak to highly developed. There are many commercial and open-source solutions available. For example, GATE, KNIME.

- LIBs:

- *Data Mining Suites(DMS):*

They focus largely on data mining. They support time series data and additional tools for text mining are available. The application is not restricted to business solutions, but it can be coupled with other functions. DMS is expensive for Example, IBM SPSS Modeler, but some open source tools exist like RapidMiner.

- Business intelligence Packages (BI's):

BI packages do not focus on data mining but they include basic data mining functionality. They have strong database coupling and are implemented via client-server architecture. They are used basically in retail stores. For example, Oracle Data Mining, DB2 Data warehouse.

- Mathematical Packages (MATs):

They have no special focus on data mining. These packages are attractive to users in algorithm development and research because they can be rapidly implemented in the form of extensions and research prototypes. For example, MATLAB and R-PLUS.

- Integration Packages (INTs):

INT's are extendables of many open-source algorithms, as stand-alone software based on JAVA. They have basic GUI support. These tools are available for different platforms, but have weak database support. These tools are attractive to users in the research field. For example, Gait-CAD, KEEL.

- EXTs:

These are small add-ons to already used tools like Excel, MATLAB, etc. their functionality is limited but very useful. The User Interface of EXTs is the same as their basic tools (Excel, MATLAB). For example, Forecaster, XLMiner.

They implement data mining methods as functions which are embedded in other tools using an Application

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

| Attribute | Focuses on |
|---|---|
| Performance | Platform variety, software architecture, heterogeneous data access, data size, efficiency, interoperability and robustness |
| Usability | Simple and efficient User Interface, user types, data visualization, error reporting, action history and domain variety |
| User Groups | Education sector, research areas, business and many other groups. Business users such as the retail store owners, use the data mining tools for Customer Relationship Management (CRM), fraud detection, inventory management, etc |
| Interaction and Visualization | The interaction between user and the data mining tool can be pure textual interface, a GUI or a graphical interface with menus |
| Import and Export of data | Ease of importing and exporting the files from and to the tools |
| Platforms | The tools that are independent of any platform are the most sought after tools |
| Licenses | Open source tools are more acceptable then licensed |
| Functionality | Algorithmic variety, prescribed methodology, model validation, data type flexibility, algorithm modifiability, data sampling, reporting and model exporting |
| Ancillary Task Support | Data cleansing, value substitution, data filtering, binning, deriving attributes, randomization, record deletion, handling blanks, metadata manipulation and result feedback |
| Data Structure | Images and video mining. The tools which support these should handle extremely large and raw datasets |
| Tasks and Methods | The tools which support supervised learning include classification, regression and fuzzy classification. The tools which support unsupervised learning will include clustering and association learning |

Table 1:
Various attributes and the area on which they focus

Programming Interface(API) for interaction between tools and and the data mining functions. There is no GUI. For example, WEKA( Java based) and MLC++ (C++ based).

- Specialties (SPECs):

These tools are similar to DMS tools, but they implement special family of methods. They are simple to use, hence used in education. For example, CART for Decision trees.

- RES:

These are the implementations of new and innovative algorithms. Any new tool or algorithm is introduced by RES implementations. They have weak Database coupling and restricted GUI. Early versions of WEKA started in this category. For example, GIFT for content based image retrieval.

- Solutions (SOLs):

These are customized group of tools for specific applications. These applications include text mining, image

Due to such wide categorization of data mining tools, giving the detailed criteria for choosing the tool becomes necessary for the user to choose an appropriate tool.

To select a data mining tool as per the user requirement, there are various characteristics that have to be kept in mind. Various methodologies had been proposed for selecting the best among the assortment of commercially available data mining software tools [4][2]. Tool selection criteria framework with the following attributes had been proposed:-

- Performance
- Functionality
- Usability
- Ancillary task support
- User Groups
- Data Structures
- Tasks and Methods
- Interaction and Visualization
- Import and Export of Data
- Platforms
- Licenses

The mentioned attributes are discussed in following table 1:

## III. USABILITY LIMITATION

Various tools for data mining are proposed such as Weka which contains various algorithms and

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

presentation of the results is done in a systematic way. Orange on the other hand consists of almost all KDD steps like data preprocessing, modeling, evaluation and visualization. Rapid miner is also a tool widely used for data mining and text miming. Many such tools are available but most of this tools are not only licensed but also require technical expertise from the user. Due to this in fact most of the companies require a technical team for data mining even though they are using a data mining tool. Having such requirement is termed as 'Usability Limitation'.

In following table usability limitation has been mentioned for existing data mining tools along with mention of whether tool is licensed or open source [6] [7] [8] [9] [10] [11] [12].

| Tool | DESCRIPTION | LICENSED | USABILITY LIMITATION |
|---|---|---|---|
| IBM SPSS Modeler | It has a visual interface that allows users to leverage data mining algorithms without programming. | Proprietary software. | Fundamental Data Mining Concepts required. |
| RapidMiner | RapidMiner uses client-server architecture with the server offered as Software as a service or on cloud infrastructures. | AGPL/Proprietary | For using RapidMiner, knowledge of programming language is must. |
| Weka | The Weka workbench contains a collection of visualization tools and algorithm for data analysis and predictive modeling. | GNU General Public License. | Algorithmic knowledge is required. |
| IPython | IPython is a command shell for interactive computing in multiple programming languages, especially in Python programming language. | BSD license | It is not supported by Windows. |
| Orange | Orange is a component based data mining tool. It includes a set of components for data pre-processing, feature scoring and filtering, modeling, model evaluation and exploration techniques. | GNU General Public License. | Can be used only by a technical user. |
| PubTator | PubTator is a web based system for assisting biocuration. It has many text mining algorithms. | Proprietary | Cannot be understood by a non technical or nin medical user. |
| SNPranker 2.0 | SNPranker 2.0 can be employed to design custom genotyping chips for disease oriented study which relies on data integration. | Proprietary. | Cannot be understood by non technical user. |

Table 2: Description of tool

## IV. PROPOSED TOOL - THE PROPHECY

The major problem faced by all major inventory based company is management of their products and inventory. However If the future sales trend is known to the company, it helps to reduce inventory and manage it more efficiently which in turn increases profit. Using data mining tools for such purpose is one on the popular choice among major inventory based companies. But for small scale or medium scale company use of licensed data mining tool or deputing technical team for this purpose annexes cost of prediction.

So in this paper we are proposing prototype of elementary prediction tool "The Prophecy-Every Data Counts!". The basic idea behind developing this tool is to provide handy reference to the inventory manager who might not possess technical expertise. This tool is proposed for the retail sector for prediction of how the company's future sales will be based on past and present

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

sales.

Our scope is to take the inventory data provided by the respective company, process the data using our prediction modeling tool and then provide the result to company that will help it in its inventory management.

The Prediction modeling techniques used Mathematical Models, Partition Model and Logical Model [13].

In this prototype, we are implementing test bench for three different types of prediction modelling.

We are using linear regression, which is a mathematical model to predict the popularity of product in retail industry. Partitioning model, in which we are using k-means clustering algorithm, gives us the most popular and the least popular cluster. In decision tree, the ID3 algorithm constructs a decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node.

These three algorithms have their merits and demerits so all the above mentioned algorithms are not capable of giving accurate prediction results individually. So we are using all the three prediction algorithms based on the nature of data and we will create testing benchmark to predict the future sales of products in retail industry.

The tool will be used in following manner:

- Various companies can provide their inventory data to us on which we will apply our prediction models
  which are:
  1. Mathematical Model
  2. Logical Model
  3. Partitioning Model
- We are thus providing the result which will help the companies in inventory management or other purposes in future.


Figure Home Page


Figure Data Entry Page

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

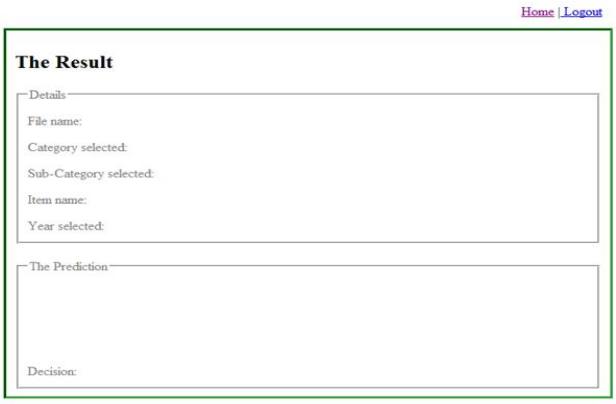


Figure Result Page

The focus of the proposed prototype is simplicity of user interface so that anyone can able to use advantages of prediction modeling.

## V. CONCLUSION

Looking into the market and various domains, the domain of retail industry is chosen for proposing a tool which can be used by an end user with little or no knowledge about data mining. The proposed tool implements three different algorithms that is Linear Regression, K-means Clustering and ID3 algorithm. The User Interface of the proposed tool is designed in a way which is easy for the user to use and also gives a rich user experience

## REFERENCES

1. Mowia Elfaki Yahia and Murtada El-mukashfi El-taher, "A new approach for evaluation of data mining techniques" IJCSI International Journal of Computer Science Issues, Vol. 7, September 2010.
2. Ralf Mikut and Markus Reischl, "Data miming tools" John Wiley & Sons, Inc, Volume 00, January/February 2011.
3. S.Hammetha Begum, "Data Mining Tools and Trends-An Overview" International Journal of Emerging Research in Management & Technology, February 2013.
4. Y.Ramamohan, K.Vasantharao, C.KalyanaChakravarti, A.S.K Ratnam, "A study of data mining tools in knowledge discovery process" International Journal of Soft computing and Engineering(IJSCE), Volume-2, Issue-3, July 2012.
5. Ogwueleka, Francisca Nonyelum and Okeke, Georgina Nkolika, "Methodology and tool selection criteria indata mining" Global Advanced Research Journal of Engineering, Technology and Innovation, Vol. 1, September 2012.
6. IBM SPSS Modeler, http://en.wikipedia.org/wiki/SPSS_Modeler
7. RapidMiner, http://en.wikipedia.org/wiki/RapidMiner
8. Weka, http://en.wikipedia.org/wiki/Weka
9. IPython, http://en.m.wikipedia.org/wiki/IPython
10. Orange, http://en.wikipedia.org/wiki/Orange
11. PubTator, http://ncbi.nlm.nih.gov/pubmed
12. SNPranker 2.0, http://ncbi.nlm.nih.gov/pubmed
13. Zbigniew Michalewicz , Martin Schmidt ,Matthew Michalewicz ,Constantin Chiriac, "Adaptive Business Intelligence", Springer | ISBN: 3540329285 | 1 edition (November 29, 2006).