Data Mining: A Review on Techniques, Tools

Mrs. Manaswini A. Parlikr[1], Mrs. G.S.Mujumdar[2], Ms. Sonali Mortale[3]
[1, 2, 3] *Lecturer, Computer Dept. Pimpri Chinchwad Polytechnic, MH, INDIA*

Abstract- **Data mining is the process of extracting the useful data, patterns and trends from a large amount of data by using techniques like clustering, classification, association and regression. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Various tools are available which supports different algorithms. A summary about data mining tools available and the supporting algorithms is the objective of this paper. Comparison between various tools has also been done to enable the users use various tools according to their requirements and applications.**

Index Terms- **Data mining, Tools, Classification, Clustering**

I. INTRODUCTION

Data mining is a part of a bigger framework, referred to as knowledge discovery in databases (KDD) that covers a complex process from data preparation to knowledge modeling.

Main data mining task is classification which has main work to assign each record of a database to one of the predefined classes. The next is clustering which works in the way that it finds groups of records instead of only one record that are close to each other according to metrics defined by user. The next task is association which defines implication rules on the basis of that subset of record attributes can be defined. Data mining is the main important step to reach the knowledge discovery. Normally for data preprocessing it goes through various process such as data cleaning, data integration, data selection and data transformation and after these it is prepared for mining task. Its main contribution is in the fields of traditional sciences as astronomy, biology, high engineering physics, medicine and investigations.

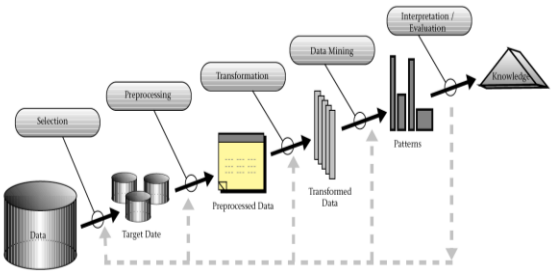

Fig 1: Data mining as KDD

Fig 2 Architecture of typical data mining system

There are a wide variety of applications in real life. Various tools are available which supports different algorithms. we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications −

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration
- Data Mining Applications

- Data mining is highly useful in the following domains −
- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

II. TECHNIQUES OF DATA MINING

A. Classification

Classification and prediction are two forms of data analysis that can be used to extract models describing the important data classes or to predict the future data trends. Such analysis can help to provide us with a better understanding of the data at large. The classification predicts categorical (discrete, unordered) labels, prediction model, and continuous valued function.

1. Neural Network

Neural Networks provide user the capabilities to select the network topology, performance parameter, learning rule and stopping criteria. An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Artificial Neural [1]

2. Decision Trees

Because of their tree structure and skill to easily generate rules the method is a favored technique for building understandable models. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

3. Genetic Algorithm

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic algorithms find application in bioinformatics, phylogenetics, computational science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields[2]

4. Rule Extraction

The taxonomy of Rule extraction contains three main criteria for evaluation of algorithms: the scope of dependency on the black box and the format of the extract description. Once rules are created and interestingness is checked they can be business where each rule performs a prediction keeping consequent as the target and the accuracy of the rule as the accuracy of the prediction which gives an opportunity for the overall system to improve and perform well.

B. Clustering

Unsupervised classification that is called as clustering or it is also known as exploratory data analysis in which there is no provision of labeled data. The main aim of clustering technique is to separate the unlabeled data set into finite and discrete set of natural and hidden data structures. There is no provision of providing accurate characterization of unobserved samples that are generated from by same probability distribution Broadly clustering has two areas based on which it can be categorized as follows:

- Hard clustering: In hard clustering same object can belong to single cluster.
- Soft clustering: In this clustering same object can belong to different clusters.

1. Clustering Process

The main four steps followed are as below:

- Feature selection or extraction: Feature selection is selecting distinguishing feature form set of candidates and extracting means which it utilizes in the transformation to generate the useful and novel features from original ones.
- Clustering algorithm design: Every clustering algorithm is affected by measures. Next is to optimize the clustering solutions
- Validation: Validations of clusters are in the sense whether the groups formed are valid or not, the data is correctly identified according to groups. These all can be checked by main three

indices which are known as testing criteria and these are as follows:

- External indices
- Internal indices
- Relative indices

- Result interpretation: Next step is to provide accuracy to user and provide a meaningful insight form original data so that efficient results can be provided.

2. Methods of clustering

There are various methods for clustering which act as a general strategy to solve the problem and to complete this, an instance of method is used called as algorithm.

two main categories - hierarchical and partitioning based methods

In hierarchical based clustering, the data sets of n elements are divided into hierarchy of groups which has tree like structure

In partitioning based methods the output is like k partitions of N dataset elements.

C. Regression

Regression is another data mining technique which is based on supervised learning and is used to predict a continuous and numerical target. It predicts number, sales, profit, square footage, temperature or mortgage rates. All these can be predicted by using regression techniques. Regression starts with data set value already known. It estimates the value by comparing already known and predicted values

1. Regression techniques or methods:

There are two types of regression techniques namely linear and non –linear.

- Linear regression: Linear regression is used where the relationship between target and predictor can be represented in straight line.
- Non- Linear Regression: In this case non linear relationship can be there and this cannot be represented as straight line.

III. TOOLS FOR DATA MINING TECHNIQUES

There are various open source tools available for data mining. Some of tools work for clustering, some for classification, regression, association and some for all.

Tool 1-Orange - Orange is a library of C++ core objects and routines that includes a large variety of standard and not-so-standard machine learning and data mining algorithms, plus routines for data input and manipulation. This includes a variety of tasks such as pretty-print of decision trees, attribute subset, bagging and boosting, and alike. Orange also includes a set of graphical widgets that use methods from core library and Orange modules. Through visual programming, widgets can be assembled together into an application by a visual programming tool called Orange Canvas[3]

Tool 2- WEKA - WEKA toolkit is a widely used toolkit for machine learning and data mining. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. WEKA contains tools for regression, classification, clustering, association rules, visualization, and data pre-processing. WEKA has become very popular with the academic and industrial researchers, and is also widely used for teaching purposes[3]

Tool 3-SCaVis - Scientific Computation and Visualization Environment. It provides environment for scientific computation, data analysis and data visualization designed for scientists, engineers and students. The program incorporates many open source software packages into a coherent interface using the concept of dynamic scripting. It provides freedom to choose a programming language, freedom to choose an operating system and freedom to share code. There is provision of multiple clipboards, multi-document support and multiple Eclipse-like bookmarks Extensive LaTeX support: a structure viewer, a build-in Bibtex manager, LaTeX equation editor and LatexTools[4]

Tool 4- Apache Mahout -The Apache Mahout™ project's goal is to build an environment for quickly creating scalable performance machine learning applications. Apache Mahout is an Open Source data Ming Tool which is designed to produce free implementations of distributed or otherwise scalable machine learning algorithms focused primarily in the areas of collaborative filtering, clustering and classification. Mahout also provides Java libraries for common mathematics operations and primitive Java collections.

Tool 5- R Software Environment - R provides free software environment for statistical computing and graphics mostly for UNIX platforms, Windows and MacOS. It is an integrated suite of software facilities like data manipulation, calculation and graphical

display. It provides a wide variety of graphical techniques as well as statistical like linear and nonlinear modeling, classical statistical tests, classification, clustering[4]

Tool 6- ML Flex – Classification algorithms developed by machine learning community. But they are implemented in diverse programming languages, have heterogeneous interfaces, and require disparate file formats. Also, because input data come in assorted formats, custom transformations often must precede classification. To address these challenges, ML-Flex, a general purpose toolbox for performing two-class and multi-class classification analyses. Via command-line interfaces, ML-Flex can invoke algorithms implemented in any programming language.[5]

Tool 7- Databionic ESOM (Emergent Self Organizing Maps) tool –

The Databionic ESOM Tools is a suite of programs to perform data mining tasks like clustering, visualization, and classification with Emergent_Self-Organizing Maps (ESOM). Features include:

Training of ESOM with different initialization methods, training algorithms, distance functions, parameter cooling strategies, ESOM grid topologies, and neighborhood kernels. Visualization of high dimensional dataspace with U-Matrix, P-Matrix, Component Planes, SDH, and more. Animated visualization of the training process. Interactive, explorative data analysis and clustering by linking ESOM to the training data, data classifications, and data descriptions. Creation of ESOM classifier and automated application to new data. Creation of non-redundant U-Maps from toroid ESOM.

Tool 8-NLTK (Natural Language Tool Kit) - NLTK provides a language processing tools, including data mining, machine learning, data capture, sentiment analysis and other language processing tasks. Because it is written in Python, you can build applications on it and customize its small tasks as well.

Tool 9-ELKI (Environment for Developing KDD-Applications Supported by Index- Structures) - ELKI (Environment for Developing KDD-Applications Supported by Index-Structures) is mainly used to cluster and find outliers. ELKI is similar to weka data mining platform, written in java, GUI graphical interface . Can be used to find outliers

Tool 10-UIMA (Unstructured Information Management Architecture) diagram – UIMA stands for "unstructured information management application". In other words, like many other data mining tools, this one seeks patterns within large data sets

Tool 11-GraphLab - GraphLab is a new parallel framework for machine learning written in C++. It is an open source project and has been designed considering the scale, variety and complexity of real world data. It incorporates various high level algorithms such as Stochastic Gradient Descent (SGD), Gradient Descent & Locking to deliver high performance experience. It helps data scientists and developers easily create and install applications at large scale. It's the presence of neat libraries for data transformation, manipulation and model visualization makes it different. In addition, it comprises of scalable machine learning toolkits which has everything (almost) required to improve machine learning models.

Tool 12-mlpy machine learning Python - It has algorithms of regression and classification. Cluster analysis can also be done for dimensionally reduction and wavelet transform. Various different algorithms like feature ranking, resampling algorithm, peak finding algorithm, error evaluation are also available. [4]

Tool 13-KEEL (Knowledge Extraction Evolutionary Learning) - KEEL (Knowledge Extraction based on Evolutionary Learning) is an open source (GPLv3) Java software tool that can be used for a large number of different knowledge data discovery tasks. KEEL provides a simple GUI based on data flow to design experiments with different datasets and computational intelligence algorithms (paying special attention to evolutionary algorithms) in order to assess the behavior of the algorithms. It contains a wide variety of classical knowledge extraction algorithms, preprocessing techniques (training set selection, feature selection, discretization, imputation methods for missing values, among others), computational intelligence based learning algorithms, hybrid models, statistical methodologies for contrasting experiments and so forth. It allows to perform a complete analysis of new computational intelligence proposals in comparison to existing ones. Moreover, KEEL has been designed with a two-fold goal: research and education

Tool 14-Scikit-learn –Scikit-Learn is a simple and efficient tool for data mining and data analysis. What is so great about it is that it's accessible to everybody, and reusable in various contexts. It is built on NumPy,SciPy, and mathplotlib. Scikit is also an open source that is commercially usable – BSD licence. Scikit-Learn has the following features:

- Classification – Identifying to which category an object belongs to
- Regression – Predicting a continuous-valued attribute associated with an object
- Clustering – Automatic grouping of similar objects into sets
- Dimensionality Reduction – Reducing the number of random variables to consider
- Model Selection – Comparing, validating and choosing parameters and models
- Preprocessing – Feature extraction and normalization

A. Comparison of various tools on different perspectives

Different factors on which categorization of tools have been stated below[4]:

Table 1 General introduction of tools:	
Tool	Aim
Orange	Visual data analysis
WEKA	General ML package
Kernlab	Kernel based classification/ Dimensionality reduction
Dlib	Portability, correctness
Nieme	Linear regression, Classification
Java-ML	Feature selection
pyML	Kernel methods
Shogun	General Purpose ML Package with particular focus on large scale learning; Kernel Methods;
Mlpy	Basic algorithms
Torch7	Neural networks
Pybrain	Reinforcement learning
Scikit-learn	General Purpose with simple API /scipy idioms

It will be also beneficial for the users to know which operation system is best suited for the data mining tool used. As there are many languages on which the tools can be used, table 3 and 4 summarizes OS and languages supported respectively.[4]

Table 2 Comparison of various tools on the basis of operating system supported:				
Tools	Linux	Windows	Mac OSX	Other Unix
Orange	Y	Y	Y	Y
WEKA	Y	Y	Y	Y
Kernlab	Y	Y	Y	Y
Dlib	Y	Y	Y	Y
Nieme	Y	Y	Y	Y
Java-ML	Y	Y	Y	Y
pyML	Y	N	Y	N
Shogun	Y	Y	Y	Y
Mlpy	Y	Y	Y	Y
Torch7	Y	Y	Y	Y
pybrain	Y	Y	N	N
Scikit-learn	Y	Y	Y	Y

Table 3 Comparison on the basis of language bindings:										
Tools	1	2	3	4	5	6	7	8	9	10
Orange	Y	N	N	N	N	N	N	N	N	N
WEKA	N	N	N	N	N	N	Y	N	N	N
Kernlab	N	Y	N	N	N	Y	N	N	N	N
Dlib	N	N	N	N	Y	N	N	N	N	N
Nieme	Y	N	N	N	Y	N	Y	N	N	N
Java-ML	N	N	N	N	N	N	Y	N	N	N
pyML	Y	N	N	N	N	N	N	N	N	N
Shogun	Y	Y	Y	Y	Y	Y	Y	N	N	N
Mlpy	Y	N	N	N	N	Y	N	N	N	N
Torch7	N	N	N	N	Y	Y	N	N	N	N
pybrain	Y	N	N	N	N	Y	N	N	N	N
Scikit-learn	Y	N	N	N	N	N	N	N	N	N

1-Python, 2-R, 3-Matlab, 4-Octave, 5- C/C++, 6-Command line, 7- Java, 8- C#, 9- Lua, 10- Ruby

Table 4 Comparison on the basis of general features:

General Features	Tools
GUI	Weka, dlib, nimene, orange,torch7,pybrain
One class classification	Shogun. weka.kernlab. dlib. pyML scikit-learn
Multi class classification	Shogun,weka,kernlab,nieme,java ml,pyML,mlpy,Pybrain,torch3,scikit-learn
Pre-processing	Pybrain,torch3,scikit learn,shogun,weka,kernlab,dlib,nieme, orange,pyML,java-ML
Regression	Pybrain,torch3,scikit-learn,shogun,weka,kernlab,dlib,nieme, orange,pyML,java-ML
Structured output learning	Shogun, nieme
Visualization	Weka,nieme,orange,pyML,mlpy,pybrain,torch3,scikit-learn
Test framework	Shogun, weka, dlib, nieme, java-ML, scikit-learn
Large scale learning	Shogun, dlib, nieme, mlpy
Semi-supervised learning	Scikit-learn
Multitask learning	Shogun
Serialization	Shogun,weka,kernlab,dlib,nieme,orange,javaml,pyML,mlpy,pybrain,scikit-learn
Image processing	Dlib

All the tools do not support all file formats. Table 6 lists six file formats and the tools which support them[4]

Table 5 Comparison of tools on the basis of file formats supported:

Tools	Binary	Arff	HDF5	CSV	Excel
Orange	N	N	N	Y	Y
WEKA	Y	Y	N	Y	N
Kernlab	N	N	Y	Y	Y
Dlib	N	N	N	N	N
Nieme	N	N	N	N	N
Java-ML	N	Y	N	Y	N
pyML	N	N	N	Y	N
Shogun	Y	N	Y	Y	N
Mlpy	N	N	N	Y	N
Torch7	N	N	N	Y	N
Pybrain	Y	N	N	N	N
Scikit-learn	Y	N	N	Y	N

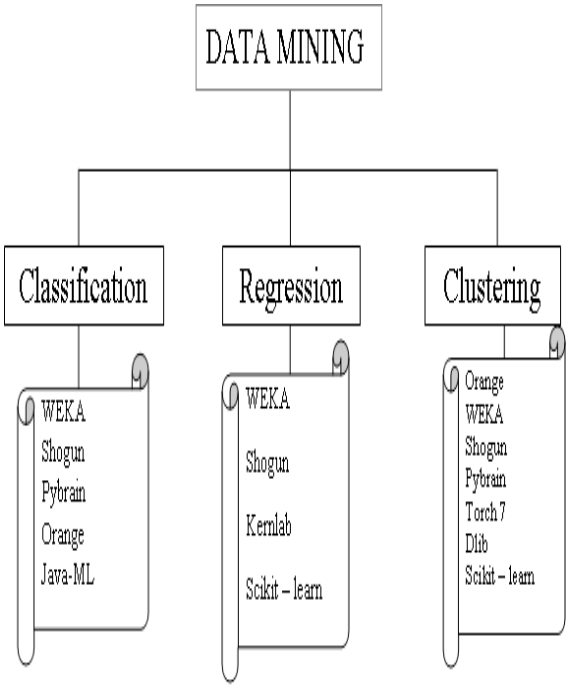

Fig. 3 Tools and Data Mining Algorithms

Major issues in data mining

1.Mining methodology and user-interaction issues

These reect the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad-hoc mining, and knowledge visualization

2.Performance issues. These include efficiency, scalability, and parallelization of data mining algorithms.

3.Issues relating to the diversity of database types.

IV. RESULT AND CONCLUSIONS

For data mining domain, the lack of explanation facilities seems to be a serious drawback as it produce opaque model, along with that accuracy is also required. To remove the deficiency of ANN and decision tree, we suggest rule extraction to produce transparent model along with accuracy. It is becoming increasingly apparent that the absence of an explanation capability in ANN systems limits the realizations of the full potential of such systems, and it is this precise deficiency that the rule extraction p Experience from the field of expert systems has shown that an explanation capability is a vital function provided by symbolic AI systems. In particular, the ability to generate even limited Craven and Shavlik in there paper listed five criteria for rule extraction, and they are as follows:

- Comprehensibility: The extent to which extracted representations are humanly comprehensible.
- Fidelity: The extent to which extracted representations accurately model the networks from which they were extracted.
- Accuracy: The ability of extracted representations to make accurate predictions on previously unseen cases.
- Scalability: The ability of the method to scale to networks with large input spaces and large numbers of weighted connections.
- Generality**:** The extent to which the method requires special training.

Data mining techniques can be widely classified into classification, regression and clustering. There are various applications of each of these. Also there are many tools available which provide methods to do different operations like WEKA, Shogun, Orange, Scikit-learn etc. The survey provided in this paper summarizes the comparison of these tools on the basis of operating system and file formats supported, general features and language bindings. This is useful for various users to select the tool best suitable for their application. All the tools do not support all the data mining operations. WEKA and Shogun supports all the three operations wiz. classification, regression and clustering while Scikit-learn supports regression and clustering operations. Orange tool supports classification and clustering. A number of applications developed by different users have been summarized which clearly shows the importance of data mining in real life.

REFERENCES

[1] dr. Yashpal singh, 2 alok singh chauhan "neural networks in data mining ",Journal of Theoretical and Applied Information Technology

[2] Gunjan Verma, Vineeta Verma," Role and Applications of Genetic Algorithm in Data Mining", International Journal of Computer Applications (0975 – 888) Volume 48– No.17, June 2012

[3] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods", (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence

[4] Mansi Gera, Shivani Goel," Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity**",** International Journal of Computer Applications (0975 – 8887) Volume 113 – No. 18, March 2015

[5] Stephen R. Piccolo, Lewis J. Frey," ML-Flex: A Flexible Toolbox for Performing Classification Analyses In Parallel", Journal of Machine Learning Research 13 (2012) 555-559

[6] PhridviRaj MSB., GuruRao CV (2013) Data mining – past, present and future – a typical survey on data streams. INTER-ENG Procedia Technology 12:255 – 263

[7] Srivastava S (2014) Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. International Journal of Computer Applications (0975 – 8887) 88:.10

[8] Soni N, Ganatra A (2012) Categorization of Several Clustering Algorithms from Different Perspective: A Review. IJARCSSE

[9] Demšar J, Zupan B (2013) Orange: Data Mining Fruitful and Fun - A Historical Perspective. Informatica 37:55–60

[10] Jain AK, Murty MN, Flynn PJ (1999) Data Clustering: A Review. ACM Computing Surveys, 31:264-323

[11] Han J, Kamber M (2001) Data Mining. Kaufmann Publishers, Morgan