International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 5 Issue 8, August-2016, Impact Factor: 1.544

Data Mining tools and techniques

Poonam Pannu

Dept. of Computer Science and Engineering, CBS Group of Institutions, MDU, Rohtak, Haryana

ABSTRACT

Today the rapid development of information technology and adoption of its several applications has created the revolution in business and various fields significantly. The growing interest in business using electronics and technology has brought vital improvement in data mining field also, since it's an important part of data accessibility. Data mining and it's applications can be viewed as one of the emerging and promising technological developments that provide efficient means to access various types of data and information available worldwide. Not only this, these applications also aids in decision making.

Keywords: Data, Data Mining, Data Mining Tools, Open Source Tools, Technical Specification, etc

1. INTRODUCTION

Data mining is an interdisciplinary subfield of science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process.

There has been a dramatic increase in amount of information and data which is stored in electronic format since last few decades. The size of data base has been in the process of continuous increment and has reached up to terabytes. This explosive rate of data increment is growing day by day and estimations tell that the amount of information in world doubles every 20 months. Thus the most important question concerned with data is its retrieval which finds the most suitable answer in data mining. Data mining is the process of extraction of predictive information from large data masses. It can also be described as a process of analyzing data from different perspectives and summarizing it into useful information. With a vast history deeply rooted in machine learning, artificial intelligence, database along with statistics data mining was coined very early.

Data mining is strongly associated with data science which involves manipulation and classification of data by applying statistical and mathematical concepts. Data mining is an important phase in knowledge discovery and includes application of discovery and analytical methods on data to produce specific models across data. Data are available everywhere. It can be used to predict the future. Usually the statistical approach is used. Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines. Due to the widespread availability of huge, complex, information-rich data sets, the ability to extract useful knowledge hidden in these data and to act on that knowledge has become increasingly important in today's competitive world .Thus data mining is analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to data owner. There are some other terms also which are related to data mining:

a) Data cleaning: to remove noise or irrelevant data.
b) Data integration: where multiple data sources may be combined.
c) Data selection: where data relevant to the analysis task are retrieved from the database.
d) Data transformation: where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance.
e) Data mining: an essential process where intelligent methods are applied in order to extract data patterns.
f) Pattern evaluation: to identify the truly interesting patterns representing knowledge based on some interestingness measures and
g) Knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

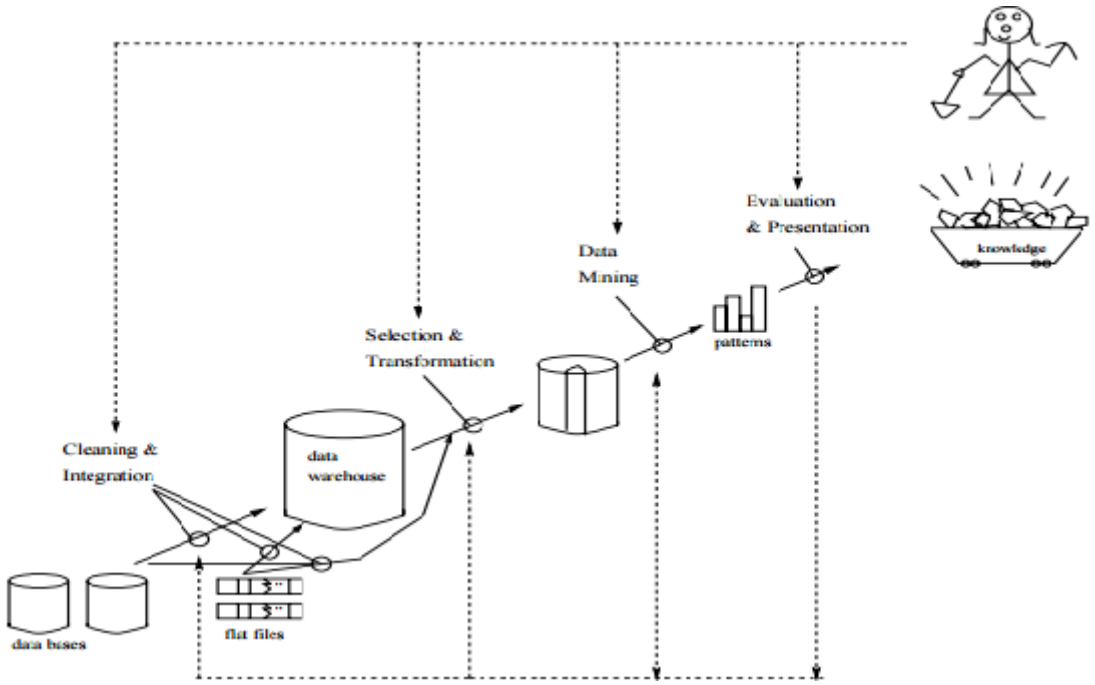

Fig 1. Data mining as a process of knowledge discovery.

Data Mining Tools

Data mining has a wide number of applications ranging from marketing and advertising of goods, services or products, artificial intelligence research, biological sciences, crime investigations to high-level government intelligence. Due to its widespread use and complexity involved in building data mining applications, a large number of Data mining tools have been developed over decades. Every tool has its own advantages and disadvantages. Within data mining, there is a group of tools that have been developed by a research community and data analysis enthusiasts; they are offered free of charge using one of the existing open-source licenses. An open-source development model usually means that the tool is a result of a community effort, not necessary supported by a single institution but instead the result of contributions from an international and informal development team. This development style offers a means of incorporating the diverse experiences Data mining provides many mining techniques to extract data from databases. Data mining tools predict future trends, behaviors, allowing business to make proactive, knowledge driven decisions. The development and application of data mining algorithms requires use of very powerful software tools. As the number of available tools continues to grow the choice of most suitable tool becomes increasingly difficult. The top six open source tools available for data mining are briefed as below.

A. WEKA (Waikato Environment for Knowledge Analysis)

Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code. The Weka (pronounced Weh-Kuh) workbench contains a collection of several tools for visualization and algorithms for analytics of data and predictive modeling, together with graphical user interfaces for easy access to this functionality.

Advantages

- It is also suitable for developing new machine learning schemes.
- Weka loads data file in formats of ARFF, CSV, and C4.5, binary. Though it is open source, Free, Extensible, Can be integrated into other java packages.

Limitation

- It lacks proper and adequate documentations and suffers from “Kitchen Sink Syndrome” where systems are updated constantly.
- Worse connectivity to Excel spreadsheet and non-Java based databases.
- CSV reader not as robust as in Rapid Miner
- Not as polished.
- Weka is much weaker in classical statistics

- Does not have the facility to save parameters for scaling to apply to future datasets.
- Does not have automatic facility for Parameter optimization of machine learning/statistical methods.

B. KEEL (Knowledge Extraction based on Evolutionary Learning)

Knowledge Extraction based on Evolutionary Learning is an application package of machine learning software tools. KEEL is designed for providing solution to data mining problems and assessing evolutionary algorithms. It has a collection of libraries for preprocessing and post-processing techniques for data manipulating, soft-computing methods in knowledge of extracting and learning, and providing scientific and research methods.

Advantages

- It includes regression, classification, clustering, and pattern mining and so on
- It contains a big collection of classical knowledge extraction algorithms, preprocessing techniques (instance selection, feature selection, discretization, imputation methods for missing values etc.), Computational Intelligence based learning algorithms, including evolutionary rule learning algorithms based on different approaches (Pittsburgh, Michigan and IRL), and hybrid models such as genetic fuzzy systems, evolutionary neural networks etc.

Limitation

Efficiency is restricted by the number of algorithms it supports as compared to other tools.

C. R (Revolution)

Revolution is a free software programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

Advantages.

- Very extensive statistical library
- It is a powerful elegant array language in the tradition of APL, Mathematica and MATLAB, but also LISP/Scheme.
- Ability to make a working machine learning program in just 40 lines of code
- Numerical programming is better integrated in R
- R has better graphics
- R is more transparent since the Orange are wrapped C++ classes.
- Easier to combine with other statistical calculations
- Import and export of data from spreadsheet is easier in R, spreadsheet are stored in a data frames that the different machine learning algorithms are operating on.
- Programming in R really is very different, you are working on a higher abstraction level, but you do lose control over the details.

Limitation:

Less specialized towards data mining.There is a steep learning curve, unless you are familiar with array languages.

KNIME (Konstanz Information Miner)

Konstanz Information Miner is an open source data analytics, reporting and integration platform. It has been used in pharmaceutical research, but is also used in other areas like CRM customer data analysis, business intelligence and financial data analysis. It is based on the Eclipse platform and, through its modular API, and is easily extensible. Custom nodes and types can be implemented in KNIME within hours thus extending KNIME to comprehend and provide first tier support for highly domain-specific data format.

Advantages

- It integrates all analysis modules of the well-known. Weka data mining environment and additional plugins allow R-scripts to be run, offering access to a vast library of statistical routines.
- It is easy to try out because it requires no installation besides downloading and unarchiving.

- The one aspect of KNIME that truly sets it apart from other data mining packages is its ability to interface with programs that allow for the visualization and analysis of molecular data

Limitations:

- Have only limited error measurement methods
- Has no wrapper methods for descriptor selection.
- Does not have automatic facility for Parameter optimization of machine learning/statistical methods.

E. RAPIDMINER

Rapidminer is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process. Rapid Miner uses a client/server model with the server offered as Software as a Service or on cloud infrastructures.

Advantages: Has the full facility for model evaluation using cross validation and independent validation sets. Over 1,500 methods for data integration, data transformation, analysis and, modeling as well as visualization – no other solution on the market offers more procedures and therefore more possibilities of defining the optimal analysis processes .Rapid Miner offers numerous procedures, especially in the area of attribute selection and for outlier detection, which no other solution offers.

Limitations: Rapid Miner is the data mining software package that is most suited for people who are accustomed to working with database files, such as in academic settings or in business settings.The reason for this is that the software requires the ability to manipulate SQL statements and files. F.

F. ORANGE

Orange is a component-based data mining and machine learning software suite, featuring a visual programming frontend for explorative data analysis and visualization, and Python bindings and libraries for scripting. It includes a set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is implemented in C++ and Python. Its graphical user interface builds upon the cross-platform framework

Advantages

It is an open source data mining package build on Python, NumPy, wrapped C, C++ and Qt. It works both as a script and with an ETL work flow GUI. It is shortest script for doing training, cross validation, algorithms comparison and prediction. It is the easiest tool to learn. Orange is written in python hence is easier for most programmers to learn It has better debugger. Scripting data mining categorization problems is simpler in Orange. Orange does not give optimum performance for association rules.

Limitations:

It is not super polished. The install is big since you need to install QT. It has limited list of machine learning algorithms Machine learning is not handled uniformly between the different libraries. Orange is weak in classical statistics; although it can compute basic statistical properties of the data, it provides no widgets for statistical testing. Reporting capabilities are limited to exporting visual representations of data models.

DATA MINING TECHNIQUES

It is main concerned with extracting useful information from large amount of databases. Data mining techniques and tools are used to find unknown patterns and trends from the data set. Its main objective is to automatically find the patterns in the dataset with minimal user effort and input. Data mining's main contribution is in decision making and in forecasting future trends of market. Many organizations use data mining as a tool these days for data analysis as it easily evaluates patterns and trends of market and produce effective results. It is the process of analyzing large sets of data and then extracting the meaning of the data. It helps in predicting future trends and patterns, allowing business in decision making. Data mining applications can answer business questions that take much time to resolve traditionally. Large amount of data which is generated for the prediction of heart disease is analyzed traditionally and is too complicated and voluminous to be processed. Data mining provides the techniques and methods for the transformation of data into useful information for decision making. These techniques make the process fast and it takes less time for the prediction system to predict the heart disease with more accuracy. In the proposed work we survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. By Applying data mining

techniques to heart disease data which needs to be processed, we can get effective results and achieve reliable performance which will help in decision making in healthcare industry. It will help the medical practitioners to diagnose the disease in less time and predict the probable complications well in advance.

a. **Association**: it is the best known and well researched method for data mining. Association is also called relation technique because patterns which are discovered from the dataset are based on the relationship between the items. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer or client behaviour.
For example, when association technique is used in heart disease prediction system, it tells us the relationship between all the attributes and sort out all the patients with all the risk factors which are required for the heart disease predictions.

b. **Classification**: it is a data mining technique which is used to classify each item in a data set into one of predefined set of classes or groups. It is a classic data mining technique which is based on machine learning. As with the most data mining solutions, classification comes with a degree of certainty. It might be probability of the object belonging to the class or it might be some other measure of how closely the object resembles other examples from that class. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in application that "given all records of employees who left the company; predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups. Goal of classification is to build a concise model that can be used to predict a class of records whose class label is not known. Example of classification is illustrated below in the following figure:

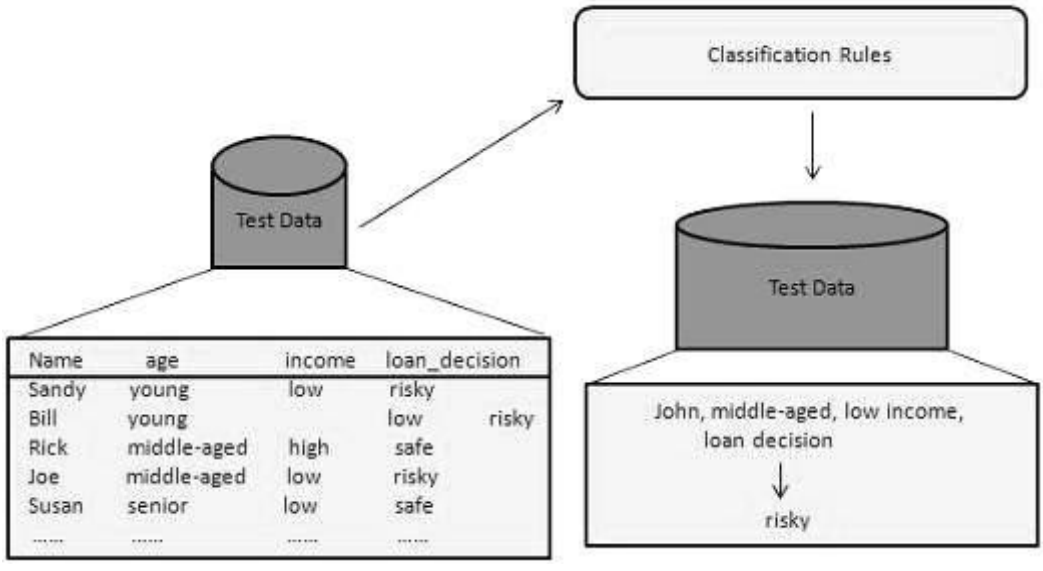

Fig. 2: Data Mining Classification

c. **Clustering**: a data mining technique that creates cluster of objects having similar characteristics is known as clustering. A cluster of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign labels to the group. Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. There is a slight difference between clustering and classification. Clustering defines classes and put objects in them while classification assigns objects into predefined classes. Clustering helps to make clusters or list of patients having same risk factor.

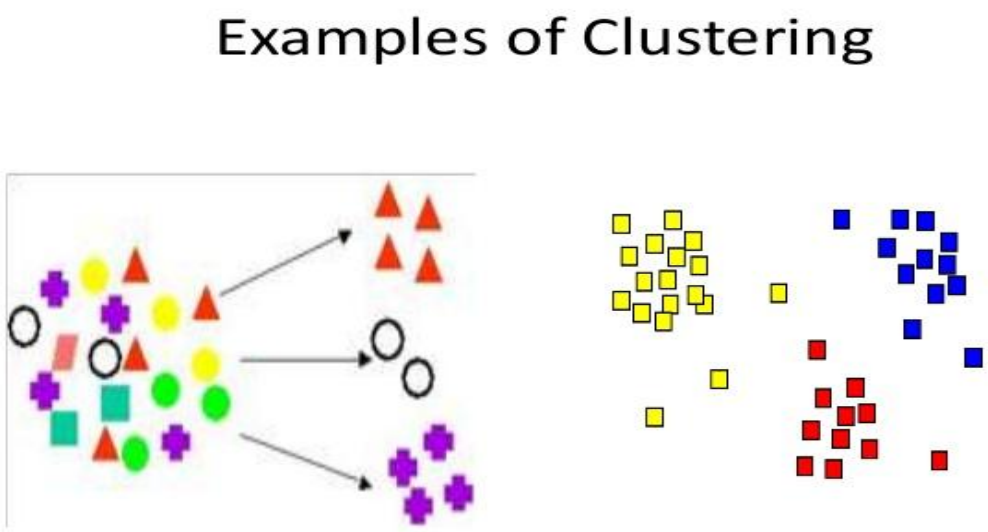

Fig. 3: Examples of Clustering

d. Neural network: it is a set of input/output units and each connection has a weight present on it. During the learning phase, network learns by adjusting the weights so as to be able to predict the correct class labels of the input tuples. Neural network have remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued input or outputs.

e. Decision tree: it is the most used data mining techniques and its model is easily understandable. The root of the decision tree is a simple question or condition that has multiple answers. Each answer leads to a set of questions or conditions which helps to determine the data so that we can take a final decision based on it.

f. Regression: Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

REFERENCES

[1]. Hand David, Mannila Heikki, Smyth Padhraic.: "Principles of data mining", Prentice hall India, pp.1, 2004.
[2]. Sethi I. K., "Layered Neural Net Design Through Decision Trees, Circuits,and Systems", IEEE International Symposium,1990.
[3]. Meheta M., Aggarwall R., Rissamen I. : "SLIQ:A fast Scalable Classifier for Data Mining", In Proc. International Conference Extending data base Technology(EDBI), Avignon, France, March 1996.
[4]. Fayyad, U., Piatesky-Shapiro, G., Smyth, P., and Uthurusamy, R. (Eds.),. Advances in Knowledge Discovery and Data Mining, AAAI Press, Cambridge, 1996..
[5]. Giorgio Barbareschi and Robbert Sanderman et al, "Socioeconomic Status and the Course of Quality of Life in Older Patients with Coronary Heart Disease", International Journal of behavioral Medicine, Vol.16, PP.197-204, 2009
[6]. HeonGyu Lee, Ki Yong Noh, and Keun Ho Ryu, "A Data Mining Approach for Coronary Heart Disease Prediction using HRV Features and Carotid Arterial Wall Thickness", International Conference on Bio Medical Engineering and Informatics, 2008.
[7]. HninWintKhaing, "Data Mining based Fragmentation and Prediction of Medical Data", IEEE, 2011.
[8]. HumarKahramanli and NovruzAllahverdi, "Design of a hybrid system for the diabetes and heart diseases", Journal of Expert Systems with Applications, Vol. 35, PP. 82–89, 2008
[9]. Indira S. FalDessai, Intelligent Heart Disease Prediction System Using Probabilistic Neural Network, International Journal on Advanced Computer Theory and Engineering, 2013.