International Journal of Scientific & Engineering Research Volume 9, Issue 4, April-2018
ISSN 2229-5518

A Review on Tools and Their Validity in Data Mining

[1]P.Karthick, [2]Mr.N.Saravanan, [3]Mr.P.Thiyagarajan

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. There are a wide variety of applications in real life. Various tools are available which supports different algorithms. A summary about data mining tools available and The supporting algorithms are 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. Different validation indices for the validation are also summarized.

Keywords—Data mining, Algorithms, Clustering and Methods.

1 Introduction

Data mining is the process of extracting useful information. Basically it is the process of discovering hidden patterns and information from the existing data. In data mining, one needs to primarily concentrate on cleansing the data so as to make it feasible for further processing. The process of cleansing the data is also called as noise limitation or noise reduction or feature elimination. This can be done by using various tools available supporting various techniques. The important consideration in data mining is whether the data to be handled static or dynamic. In general, static data is easy to handle as it is known earlier and stored. Dynamic data refers to high voluminous and continuously changing information which is not stored earlier for analyzing and processing like static data. It is difficult to maintain dynamic data as it changes with time. Many algorithms are used to analyze the data of interest. Data can be sequential, audio signal, video signal, spatio-temporal, temporal, time series, etc.

Data mining is a part of a bigger framework, referred to as knowledge discovery in data bases (KDD) that covers a complex process from data preparation to knowledge modeling [2]. 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. Then exit 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 the see it is prepared for mining task.

2 Techniques of Data Mining

To analyze large amount of data, data mining came into picture and is also called as KDD process. To complete this process various techniques developed so far are explained in this section. KDD is the overall process which is shown in figure 1:

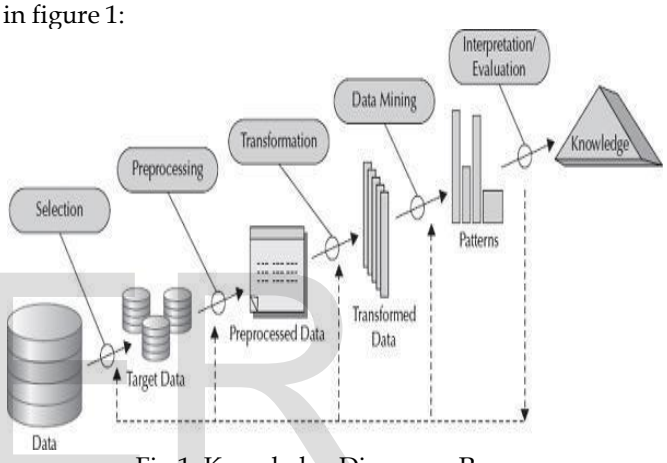

Fig 1: Knowledge Discovery Process

Data mining is the core part of the knowledge discovery process. In this, process may consist of the following steps Data selection, Data cleaning, Data transformation, pattern searching (data mining),finding presentation, finding interpretation and finding evaluation. The data mining and KDD often used interchangeably because Data mining is the key part of KDD process. The term Knowledge Discovery in Databases or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods The unifying goal of the KDD process is to extract knowledge from data in the context of large data bases. It does this by using data mining methods(algorithms) to extract(identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, sub sampling, and transformations of that database.

3 Classification

Classification is one of the data mining techniques which are useful for predicting group membership for data instances. Classification is a supervised kind of machine learning in which the provision of labeled data in advance. By providing training the data can be trained and we can predict the future of data. Prediction is in the form of predicting the class to which data can belong. Training is based on the training sample provided. Basically there are

- [1]P.Karthick, II-Year MCA, Priyadarshini Engineering College, Vaniyambadi, Email: karthickmonitpt26@gmail.com.
- [2]Mr.N.Saravanan, Asst.Prof MCA, Priyadarshini Engineering College, Vaniyambadi, E-Mail: saran07@gmail.com.
- [3]Mr.P.Thiyagarajan, Asst.Prof MCA, Priyadarshini Engineering College, Vaniyambadi, Email: ajay.thiyagaraj@gmail.com.

two types of attributes available that are output or dependent attribute and input or the independent attribute [9]. In the supervised classification, there is mapping of input data set to finite set of discrete class labels. Input data set X€ Ri, where I is the input space dimensionally and discrete class label Y € 1.....T, where T is the total number of class types. And this is
Modeled in the term of equation Y=Y(x, w), w is the vector of adjustable parameters.

4 CLASSIFICATION METHODS IN DATA MINING ARE AS FOLLOWS

Decision tree induction

From the class labeled tuples Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The top most decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

4.1 Rule–based classification

It is represented by set of IF-THEN rules. First of all how many of the rules are examined and next care is about how the rules are build and can be generated from decision tree or it may be generated from training data using sequential covering algorithm. Expression for rule is:
IF condition THEN conclusion
Now we define accuracy and coverage of S by following Expression [10]
Coverage (R) = ������ I◘I
Coverage (R) = ��������/ ������

4.2 Classification by back propagation

Back propagation is a neural network learning algorithm. Neural network learning is often called connectionist learning as it builds connections. It is feasible for that application where long times training is required. The most popular neural network algorithm is back propagation. This algorithm proceeds in the way that it iteratively performs processing of data and it learns by comparing the results with the target value given earlier.

4.3 Lazy learners

Eager learner's the form in which generalization model is being developed earlier before new tuple is being received for classifying. In lazy learner approach when given a training tuple it simply stores it and waits until a test tuple is given. It supports incremental learning. Some of the examples of lazy learner are K-nearest neighbor classifier and case- based reasoning classifiers [11].

5 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 dataset into finite and Discrete set of natural and hidden data structures. There is no provision of providing accurate characterization of un building groups. So to improve this new concept come into

6 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. Broadly clustering methods can be divided into two main categories which have number of instances. On the basis of that we have hierarchical and partitioning based methods. In hierarchical based clustering, the datasets 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.

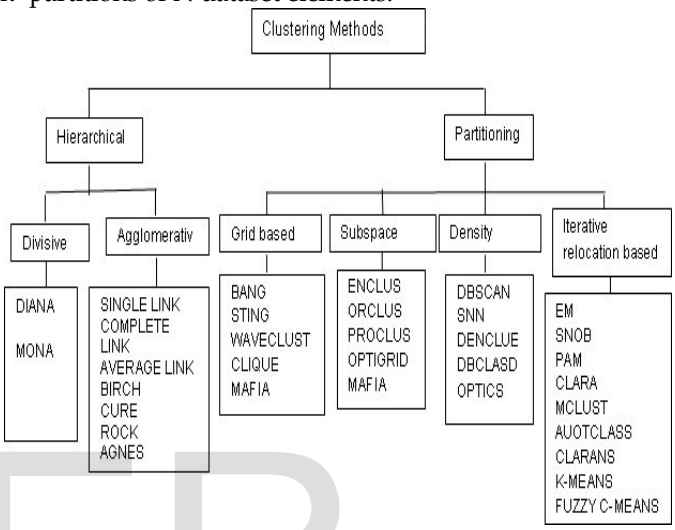

Fig 3: Categorization of clustering methods and algorithms

7 RELATED WORK ON APPLICATIONS OF DATA MINING TECHNIQUES

Data mining techniques are used in many applications. The Effect and future trends have been stated. Many users have designed prediction systems using these techniques. There is a study of various factors that affect academic performance and for that the data of pharmacy students have taken focusing on which will help students to improve their performance [21] A paper by Kriegeletal [22]. It focuses on building the classification model to predict the performance of employees. Many factors have been included and on the basis of that the experiment has done by Radaidehand Nagi [23]. Another paper by SudhaandVijiyarani [24] is on the prediction of diseases as heart diseases, diabetics etc. by using data mining techniques. By using classification techniques like decision tree, naïve bayes a prediction model is designed [25].
Use of K-means algorithm is very useful in designing many applications. Extension of K-means algorithm can be done to improve the performance [26] in a paper by Ngaietal. A review of the classification scheme for the application of financial fraud detection using data mining technique is done [27]. A survey by Andréetal. Shows different perspectives that in the data obtained by partitioning done by clustering ensembles, data can be improved by applying more steps and this all could be done through genetic programming approach [28]. As in unsupervised learning, there is no target attribute known in advance and there may be some time no comparison and correction in

picture that is bounded rationality to reveal feature saliency in clustering problem designed [29].

8 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. There are various algorithms for each technique asdiscussedinsection2.This section describes features of different tools and which tools can be used to implement which algorithm.

9 FEATURES OF DIFFERENT TOOLS

9.1 WEKA

WEKA stands for Waika to Environment for Knowledge Analysis. It is developed in Java programming language. It contains tools for data preprocessing, classification, clustering, association rules and visualization. It is not capable for multi-relational data mining. Data file can be used in any format like ARFF (attribute relation file format), CSV (comma separated values), and C4.5and binary and can be read form a URL or from SQL data base as well by using JDBC. One additional feature is that data sources, classifier set care called as beans and these can be connected graphically [2].

9.2 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 in corporate, there are 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 La Te X support: a structure viewer, a build-in Bibte x manager, La Te X equation editor and Latex Tools [42, 43]

9.3 Apache Mahout

Its goal is to build machine learning library scalable to large data set. For Classification following algorithms are included: Logistic Regression, Naive Bayes/ Complementary Naive Bayes, Random Forest, Hidden Markov Models, Multilayer Perceptron. For Clustering following algorithms are included: Canopy Clustering, k-Means Clustering, Fuzzy k-Means, Streaming k-Means, Spectral Clustering by Sean Owen and Sebastian Schelter [44].

9.4 R Software Environment

R provides free software environment for statistical Computing and graphics mostly for UNIX platforms, Windows and Mac-OS. 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 non-linear modeling, classical statistical tests, classification, clustering [10].

9.5 MLFlex

ML uses machine learning algorithms to derive models from independent variables with the purpose of predicting the values of a dependent (class) variable.

9.6 Databionic ESOM (Emergent Self Organizing Maps) tool

On can do Preprocessing, Training, Visualization, Data analysis, Clustering, Projection, and Classification using this tool. Training data is set of points from a high dimensional space called data space. The two most common training algorithms are online and batch training. Both of these training algorithms will search the closest prototype for each data point that is best match. Online training, there is immediately update of best match but in batch training all the best matches are being collected and then update if performed collectively [10].

9.7 NLTK (Natural Language Tool Kit)

NLTK is a leading platform for building Python programs to work with human language data. Independence of file parsers or database connections, data types, distances, distances functions, and data mining algorithms [45].

9.8 UIMA (Unstructured Information Management Architecture) diagram

Large amount of unstructured information can be analyzed to get relevant information. It enables application to be decomposed into components. Working of framework is to manage these Components and flow between them. Basic availability is frameworks, components and infrastructure [46, 47].

9.9 GraphLab

Graph Lab has several algorithms already implemented in its toolkit. One can also implement one's own algorithm on top of our graph programming API [48].

9.10 Scikit-learn

Scikit-learn are also a free package. It is in Python which extends the functionality of Num Pyand SciPy packages. It also uses them at plot lib package for plotting charts.

The package supports most of the core DM algorithms except including classification rules and association rules.

10 CONCLUSIONS

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.

REFERENCES

[1] Mrs. Bharati M. Ramageri, "Data Mining Techniques and Applications," Indian Journal of Computer Science and Engineering, Vol. 1 No. 4, pp. 301-305

[2] Hemlata Sahu, Shalini Shrma and Seema Gondhalakar, "A Brief Overview on Data Mining Survey," International Journal of Computer Technology and Electronics Engineering (IJCTEE)., Vol.1, Issue 3, pp.114-121

[3] Kalyani M Raval, "Data Mining Techniques," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2Issue 10,pp.439-442.

[4] Sangeeta Goele, Nisha Chanana, "Data Mining Trend In Past, Current And Future," International Journal of Computing & Business Research, in Proc.I-Society 2012, 2012

[5] Mr. S. P. Deshpande and Dr. V. M. Thakare, "Data Mining System and Applications: A Review, "International Journal of Distributed and Parallel systems (IJDPS) Vol.1, No.1, September 2010, pp.32-44

[6] Y. Ramamohan, K. Vasantharao, C. Kalyana Chakravarti, and A.S.K.Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process,"International Journal of Soft Computing and Engineering (IJSCE), Vol. 2, Issue-3, July 2012, pp.19 1-1994

[7] Kleinberg J (2002) An impossibility theorem for clustering. Conf. Advances in Neural Information Processing Systems, 15:463–470