Data Mining Tools and Techniques: A review

Prof. Vaishali Jabade
dept. of Electronics and Telecommunications Engineering,
Vishwakarma Institute of Technology,
Pune, India.

Abhijeet Jadhav
dept. of Electronics and Telecommunications Engineering,
Vishwakarma Institute of Technology,
Pune, India

Abstract: *Over the past few years, data mining has made a great advancement still the main problem of missing data has remained a dispute for different types of Data Mining Algorithms and Tools. Data Mining is an activity of extracting some useful data from large datasets/databases. The mostly used techniques in data mining are classification, clustering, regression, association, etc. In real life, there are many applications which works on different algorithms with the help of different types of tools. The objective of this paper is to provide a basic summary on different types of tools available for data mining and different types of algorithm based on it. Discussion of various tools and algorithms has done in this paper which will enable users to access different tools according to their use and applications.*

Keywords: ***Data Mining, clustering, classification, database, dataset, data***

1. INTRODUCTION

Data mining is a process of extracting useful data from a bunch of large data. Data mining is a common term used for knowledge Discovery from Data, or KDD. KDD is a process of discovering and extracting the hidden patterns and useful information from data. While mining the data, first of all, data cleansing should be done to make data more feasible for future processing. In the process of data cleansing, noise reduction or noise elimination or feature elimination is done and it can be done using different tools and by using various techniques. There are basically two types of data namely Static and Dynamic which will be another important consideration for data mining. Static data is known and stored whereas Dynamic data is high voluminous, is changing continuously and is not stored earlier to analyze and process. So, handling of data in case of Static data is easy as compare to dynamic data. Dynamic data is difficult to maintain as it changes with respect to time. There are a lot of algorithms used for analyzing the data of interest. Different techniques are used for different data types. Data can be sequential data, audio data signal, video data signal, spatio-temporal data, temporal data, time series data, etc.

2. KDD

Primarily, raw data needs to be converted to some form from where the data can be processed. Knowledge Discovery in Database is a process of a few steps which will lead raw data collections to some form of new knowledge. This process mainly includes the steps shown in fig. (1) below:

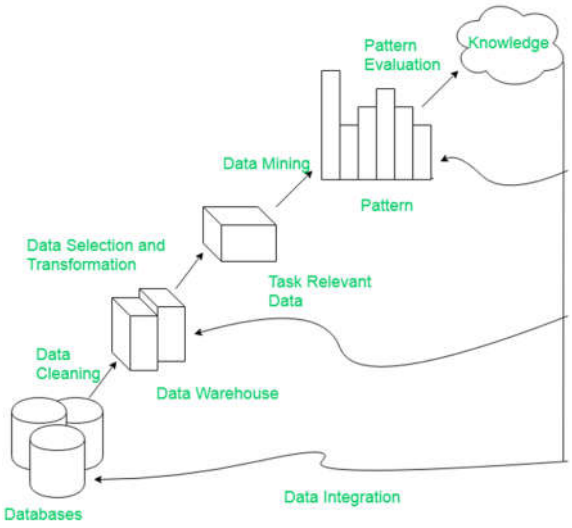

Fig. (1) – steps in KDD

Data cleansing: It is also known as the data cleaning. In this step, the noise from the data is removed and irrelevant data is also removed.

Data integration: Data integration is a step where different data types are combined. In this, multiple heterogeneous data sources may be combined in a common source.

Data selection: In this stage, the data relevant to the analysis is decided and retrieved from the database.

Data transformation: At this stage, data is consolidated or summarized. The selected data s transformed into the forms which appropriate for the mining.

Data mining: The most important and essential process where various techniques are applied to extract patterns from the selected data.

Pattern evaluation: In this stage, patterns representing knowledge are identified based on given the measures given.

Knowledge representation: The final phase in KDD in which the knowledge discovered is visually represented to the user. This stage uses different types of visualization techniques to help understand and interpret the results to the users.

3. DATA MINING TECHNIQUES

In this paper, we have studied four data mining techniques along with different tools that are used for data mining. The techniques include:

1. Classification
2. Clustering
3. Regression

1. Classification

Classification is an approach based on supervised machine learning. In supervised learning, advantage of labeled data in advance is there. It is a process which contains two steps. In the first step, a model is trained by providing training data and in second step we can predict the future for specific data. Training is based on the sample data provided and prediction is in the form of specific class to which the given data belong. There are basically two attributes available which are input and output. In this process, mapping of input data set to discrete class labels s done. Input data set, let's say X € Ri, where 'i' is the input space and the discrete class labels 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.

There are different types of classification techniques in data mining. Some of them are mentioned below:

- Decision Tree Induction:

Decision tree induction is basically a structure with three components. Three components are internal node, branches and leaf nodes. The topmost node in the structure is a root node. Each internal node in this structure denotes a test on an attribute. Each branch represents the outcome of the test and the leaf node is used to denote class label. It is a two-step process consists of learning and testing. The main goal for this structure is to predict the output for continuous attributes but this approach can be less appropriate for estimating such a tasks. There can be errors in predicting classes by using decision trees. Also, the costs of pruning algorithms are more and building decision tree is also an expensive task because at each level splitting of node is there.

The basic structure for a Decision Tree Induction is shown in fig. (2) below:

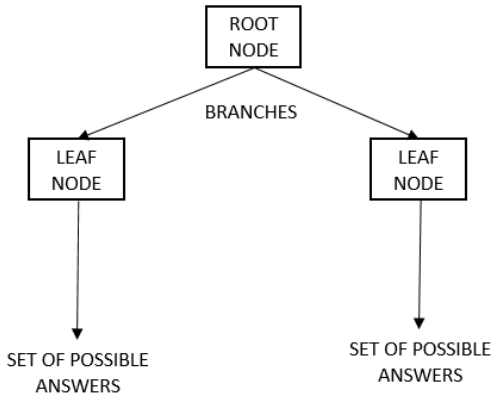

Fig. (2)-Decision Tree Induction

- Rule Based Technique:

Rule-based technique is a popular class of techniques mostly used in machine learning and data mining. They both have the goal to find regularities in data which can be expressed in the form of IF-THEN. These type of classifiers use IF-THEN rules for classification and can be expressed as:

IF condition THEN conclusion

IF part of the rule is known as precondition and THEN part of the rule is known as rule consequent. The precondition part consists of one or more attribute tests and are logically separated by AND. The consequent

part is a prediction class. Primarily, number of these rules are examined and next part is about how these rules are build and can be created by the use of decision tree or it can be generated from training data using sequential covering algorithm.

The accuracy and Coverage are defined by following expressions:

Coverage (R) = $Ntotal/$I$\boldsymbol{D}$I

Coverage (R) = $Ncorrect/Ntotal$

- Classification by Backpropagation:

Backpropagation is a technique which uses neural network learning algorithm. It was Started by psychologists and neurobiologists for the development and testing of the computational analogues of neurons. Neural network learning builds networks internally and is called as connectionist learning. It is viable to the applications where long times training is needed. One of the most popular neural network algorithm for data mining is backpropagation. This algorithm proceeds in a way that the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples by the comparison of the results with the target value given earlier.

- Lazy Learners:

There are basically two learners used in machine learning and data mining namely eager learners and lazy learners. Eager learning is a method where generalization of the training data is done before receiving queries whereas, in lazy learners, generalization of the training data is delayed until a query is made to the system.

Examples of lazy learners include K-nearest neighbor classifier and case- based reasoning classifiers.

2. Clustering

Clustering is basically an Unsupervised classification method or an as exploratory data analysis as it has no provision for labeled data. The main goal of this technique is to separate the unlabeled data into finite and discrete set of natural and hidden data structures. Clustering uses input data which determines patterns, anomalies, or any other similarities in its input data. A clustering algorithm is said to be good if it obtains intra-cluster similarities high and inter-cluster similarity low.

Clustering can be subdivided into categories based on prediction of clusters:

- Hard Clustering: one object can belong to a single cluster.
- Soft Clustering: one object can belong to different clusters.

Let's say there is set of input patterns given, Y= {y1,.....,yi,.....yN}, where yi= (yi1,.....yid) T ∈ R^d and each is yjd known as variable, feature, dimension or attribute.

Hard partitioning results are like:

i. C= {C1,.....,CK} where (K≤N) and
ii. Ci ≠ ϕ, i=1, 2,...., N
iii. ∪i=1K $Ci=Y$
iv. $Ci \cap Cj=\emptyset, i,j=1,2,\ldots..K \text{ and } i \neq j$

Whereas, Hierarchical clustering has a different approach of representing the output that is tree like structure, partition of Y,P= $P1,\ldots\ldots Pr$ *where* $r \leq N$ *and* $Ci \in Pl$ *and* $Cj \in Pm$ and l > m imply $Ci \in Cj$ *for all* $i,j \neq i,l$ m=1,2,….r

- Process of Clustering:

The clustering process is a step by step process in which the results can be verified and which includes various steps. There are mainly four steps and are as follows:

Feature selection and extraction: Feature selection is a process of selecting differentiating feature form set of candidates and extracting means which it uses in the transformation to generate the useful and novel features from original ones.

Clustering algorithm design: Next process is to select a design of clustering algorithm. Each and every clustering algorithm is affected by the measures. Next part is to optimize the clustering solutions.

Validation: Next process is validation means to check if the groups formed are valid or not. In this, the data is correctly identified with respect to groups. To check validation, it is tested by three main indices which is called testing criteria. The three indices are: External Indices, Internal Indices and Relative Indices. These indices are defined on various clustering structures like hierarchal clustering, partitioning clustering, etc.

Result interpretation: The last part of this procedure is to provide accuracy, meaningful insights and result interpretation to the users.

- Clustering Methods:

There are different methods for clustering which act as a strategy for solving problems. The methods that are used for this are called algorithms. There are mainly two important types of clustering which have number of instances. On this basis, hierarchical and partitioning based methods are the types manly used. Hierarchical based clustering divides the data sets of n elements into hierarchy of groups which has tree like structure whereas, partitioning based method, the output is like k partitions of N dataset elements.

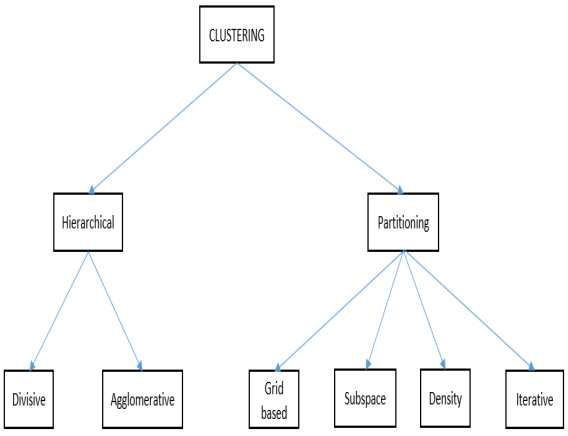

Fig. (3)- Clustering methods

Hierarchical Methods: In this method, a hierarchical decomposition of the given set of data objects is created which is a tree like structure. This method is further classified into two categories. One approach is divisive and another approach is agglomerative. Examples of hierarchical include BIRCH, CURE, ROCK, Chameleon.

Agglomerative approach is known as bottom-up approach. It is started with each object forming a separate group and it keeps on merging the objects/groups which are closer. Until all of the groups are merged into one or until the termination condition holds, it keeps on repeating.

Divisive Approach is known as top-down approach. It is completely opposite to agglomerative approach. It starts with all objects in single cluster and then n continuous iteration, it starts splitting the cluster into smaller ones. Until there is each object in one cluster or the termination condition holds it keeps on repeating it.

Partitioning methods: In this method, the dataset is simply partitioned into n objects. k number of partitions known as clusters are done with n objects such that $k \leq n$. Different types of approaches are also there in partitioning methods. Examples of Partitioning include FCM, K-means, PAM CLARA, CLARANS

Grid-based Method is a method where objects form a grid and each step this structure is followed. The object space in this is quantized into number of cells which forms a grid. The main advantage of this method is the time. It has very high processing time. Examples are OptiGrid, CLIQUE, STING.

Subspace based method use subspace of actual document and its main aim is to work with high dimensional data.

In density based approach, the given cluster is increase to cover the neighborhood exceeds some threshold value. Examples include DBSCAN, DBCLASD, OPTICS, DENCLUE.

Relocation based methods are used like a point of view in which it identifies the unknown parameters of the clusters.

3. Regression

Another technique in data mining is regression which is based on supervised learning and is usually used to predict a numeric, continuous data. It is used predict number, profit, sales, square footage, temperature, mortgage rates, etc. All above mentioned can be predicted by using regression techniques. It is a process which starts with data set value already known and train it. It estimates the output by comparing the values which are already known and predicted values. These values are summarized in a model. Main goal from this is to reduce error and to give accurate value to the result.

- Regression Methods:

There are basically two types of regression methods are used widely.

Linear Regression: It is used when the relationship between target and predictor can be represented in a straight line.

Non-linear Regression: In this, non-linear relationship can be there between predictor and target and may not be represented by a straight line.

4. TOOLS FOR DATA MINING

There are different types open source tools available for data mining processes. Different tools work on different methods like classification, regression, clustering or association and some them also works on more than one methods. Also, there are various algorithms available for each methods as discussed earlier. In this section, an overview for some tools used for data mining is provided.

1. Orange:

It is one of the open source tool used for data visualization and analysis. In Orange, data mining is done with the help of visual programming and Python Scripting. In this, regressions methods are also used.

2. WEKA:

Waikato Environment for Knowledge Analysis (WEKA) is developed in Java. It contains almost tools for almost all the processes like data visualization, preprocessing, clustering, classification, association, etc. In this, data files can be in any format like CSV, SQL using JDBC, ARFF, etc. the only disadvantage is that it cannot work for multi relational data mining. The special feature is that the entities like classifiers can be connected graphically.

3. SCaVis:

SCaVis stands for Scientific Computation and Visualization Environment. It is designed to provide platform for different types of data visualization, data analysis and scientific computation. This tool has a lot of open source software packages into an interface with dynamic scripting. It can provide different operating systems and different programming languages.

4. Apache Mahout:

It is used to build machine learning library more scalable to large data set. It supports both classification as well as clustering methods for data mining. Naive Bayes, Logistic Regression, Random Forest, Hidden Markov Models, etc. are the classification algorithms are included in it. k-Means Clustering, Fuzzy k-Means, Streaming k-Means, etc. are the clustering algorithms are included in it.

5. R Software Environment:

R works mostly on Windows, MacOS and UNIX platforms. It basically provides free software environment for statistical computing and graphics. It is a combined tool which provides different facilities like data manipulation, calculations, statistics and graphs. It has a set of wide varieties of graphical techniques and statistical techniques like modeling, statistical tests, classification, clustering, etc.

6. Scikit-learn:

Another free package is Scikit-learn. It works in Python and extends the functionalities of Python packages like NumPy,SciPy and matplotlib. Except classification rules and association rules, this package supports most Data Mning algorithms.

7. GraphLab:

GraphLab is a tool where several algorithms are already implemented in its toolkit. We can also implement our own algorithm on top of our graph API.

8. mlpy machine learning Python:

It consists of algorithms of regression and classification. For dimension reduction or wavelet transform, cluster analysis can also be done in it. Number of algorithms including peak finding, feature ranking, error evaluation, etc. are also available.

9. KEEL:

Knowledge Extraction Evolutionary Learning is also an open source tool. It is based on Java (GPLv3). It provides users with the access of behavior of evolutionary learning and basic soft computing based techniques for different kinds of data mining issues needs to be handled.

10. Databionic ESOM:

Databionic Emergent Self Organizing Maps is a tool where one can do many processes. It provides preprocessing, training, data visualization, data analysis, clustering, projection and Classification. In this, there are two main training algorithms are there namely online training and batch training. Both of them will search for closest prototype. In online training, there will be immediate update whereas in batch training the best matches are collected and then updated.

So, we have described ten tools that one can use for data mining processes which supports different data mining techniques. In the following table, we have shown the different data mining techniques which can be implemented in various tools.

DATA MINING TECHNIQUES->	CLASSIFCATION	CLUSTERING	REGRESSION
TOOLS USED->	• WEKA • Shogun • Pybrain • Orange • Java-ML	• Orange • Dlib • WEKA • Shogun • Pybrain • Torch7 • Scikit-learn	• Shogun • WEKA • Scikit-learn • Kernlab

Table (1)

5. CONCLUSION

In this paper, a review on different types of data mining techniques, algorithms and tools used is provided. Data mining is a process extracting useful information from data which is stored in datasets or databases. Different techniques along different types of algorithms for each technique and tools used for them is specified in this paper. There are tools like WEKA and shogun which supports most of the techniques for data mining which includes classification, clustering and regressions. According to the use, a user can choose any of the tools and techniques for his applications.

6. REFRENCES

[1] Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”,Second Edition 2006.

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

[3]Han. J, Kamber. M, Pei. J, “ Data Mining Concepts and Techniques”, Third edition The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011

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

[5] Demšar J, Zupan B (2013) Orange: Data Mining Fruitful and Fun - A Historical Perspective. Informatica.

[6] Gupta GK (2012) Introduction to data mining with case studies PHI, New Delhi

[7] Kumar R, Kapil AK, Bhatia (2012) A Modified tree classification in data mining. Global Journals Inc. 12, 12: 58-63

[8] Zhao Q, Fränti P (2014) WB-index: A sum-of-squares based index for cluster validity. Data & Knowledge Engineering 92:77–89

[9] Tayel , Salma, et al. “Rule-based Complaint Detection using RapidMiner”, Conference: RCOMM 2013, At Porto, Portugal, Volume: 141-149,2014

[11] Velmurugan T (2014) Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data. Applied Soft Computing.

[12] P. Subathra, R. Deepika, K. Yamini, P. Arunprasad, S.k Vasudevan, A Study of Open Source DataMining Tools and its Applications, 10(10), 2015

[13] André L.V. Coelho, , Everlândio Fernandes, Katti Faceli (2011) Multi-objective design of hierarchical consensus functions for clustering ensembles via genetic programming Decision Support Systems

[14] Aviad B, Roy G (2012) A decision support method, based on bounded rationality concepts, to reveal feature saliency in clustering problems. Decision Support Systems 54: 292–303

[15] Combes C, Azema J (2013) Clustering using principal component analysis applied to Autonomy – disability of elderly people. Decision Support Systems 55:578–586

[16] Sandeep, Priyanka, Bansal R (2014) Performance Comparison of Various Partition based Clustering Algorithms. IJEMR pp. 216-223

[17] Rao GN, Ramachandra M (2014) A Study on the Academic Performance of the Students by Applying K-Means Algorithm. IJETCAS 14-180

[18] Lin PL, Po-Huang PW ,Kuo PH , Lai YH (2014) A size-insensitive integrity-based fuzzy c-means method for data clustering. Pattern Recognition

[19] Jacques J, Preda C (2014) Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis 71:92–106

[20] Padmaja S and Fatima SS (2013) Opinion Mining and Sentiment Analysis –An Assessment of People's Belief: A Survey. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) 4(1)