Data Mining: You’ve missed it If Not Used

Kanchan A. Khedikar[1] ,Mr. L.M.R.J.Lobo[2]
[1]*(Student M.E.C.S.E. Walchand Institute of Technology, Solapur)*
[2]*(Professor & Head,Department Information Technology, Walchand Institute of technology, Solapur)*

Abstract: *Mining data is related to extracting interesting patterns or knowledge from huge amount of data available on existing resources. By interesting we mean, patterns are non- trivial implicit, previously unknown and potentially useful facts. In Data Mining techniques, huge amounts of data is being mined. The goal of data mining is to convert such data into useful information & knowledge. This paper relates about data mining models, its applications in various fields and tools used for data mining. This survey paper gives explanation of different data mining techniques such as clustering, classification, association rule. Various tasks like Dependency analysis, Class identification, Concept description, Deviation detection and Data visualization are also touched. This paper also explains two very important data mining tools that is Weka and Orange.*

Key Words- *Data mining, Knowledge discovery database, classification, clustering, association rules, web mining*

I. Introduction

It becomes virtually impossible for individuals or groups with limited resources to find and gain any insight from the data. Data Mining uses tools and techniques for the ‘extraction’ or mining’knowledge from large amounts of data [1]. It is about finding patterns and relationships within data that can possibly result in new knowledge. These relationships can also result in predictors of future outcomes. Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining is ready for application in the business community because it is supported by three technologies that are Massive data collection, Powerful multiprocessor computers and Data mining algorithms.

The importance of data mining has been established for business applications, criminal investigations, bio-medicine and more recently counter-terrorism or fraud detection. Data mining is also known as Knowledge Discovery in Databases (KDD) [2] (See Fig.1)

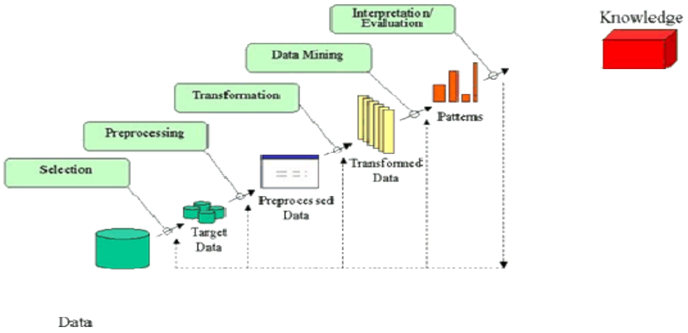


Fig 1: KDD process [3]

Steps involved in KDD Process are Data selection, Data cleaning, Data integration, Data transformation, Data mining, Pattern evaluation and Knowledge presentation. Data selection deals with the selection of data from large amount of databases, Data cleaning is used to remove noise and inconsistent data and in data integration multiple data sources may be combined which is also known as data pre-processing. Data transformation is transforming data into the form, which is appropriate for mining by performing summary or aggregation operation. Data mining is an essential process where intelligent methods (such as various algorithms) are applied in order to extract data patterns. Pattern evaluation is used to identify the truly interesting patterns representing knowledge based on some interestingness measures and in knowledge presentation visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data mining uses a broad family of computational methods that include statistical analysis, decision trees, neural networks, rule induction and refinement, and graphic visualization [3]. Data mining supports different techniques like classification, clustering, association rule.

II. Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as Decision Trees (DT's), Support Vector Machine (SVM), Genetic Algorithms (GAs) / Evolutionary Programming (EP), Fuzzy Sets, Neural Networks, Rough Sets [4]. For example, we can apply classification in application that "given all past records of employees who left the company, predict which current employees are probably to leave in the future." In this case, we divide the employee's records into two groups that are "leave" and "stay". And then we can ask our data mining software to classify the employees into each group. Popular classification techniques include decision trees and neural networks [5]. Another most important example of classification is in the area of sports.

In sports they mine sports video annotation data to extract knowledge about match play sequences and applying that knowledge for classification of players for developing player specific training taxonomy. To analyze individual player's performances a classification technique is used to classify them into appropriate groups using the frequently played patterns and other performance indices like strike rate, six-runs and four runs. This classification helps the coaches to know the current form of the player and to understand their strengths and weaknesses. With this information, a coach can assess the effectiveness of certain coaching decisions and formulate game strategy for subsequent games. Classification mechanism is applied to analyze each and every individual player's strengths and weaknesses to fix them into a respective class of training taxonomy.

III. Clustering

Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique.

Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes. To make the concept clearer, we can take library as an example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without hassle. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library. Clustering is a method in which we make a cluster of objects that are similar in characteristics. It is a technique in which, the information is logically similar and physically stored together. See figure 2.

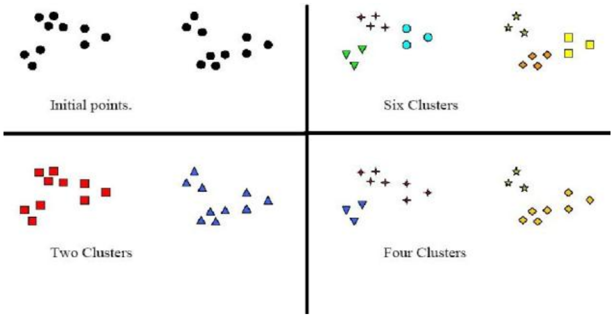


Fig 2 : Grouping of object in clusters[3]

Types of clustering are Partitioned Clustering and Hierarchical clustering, Partitional clustering is a division of data objects into non- overlapping subsets (clusters) such that each data object is in exactly one subset. Hierarchical clustering is a set of nested clusters organized as a hierarchical tree [6]. There are several types of Hierarchical and Non- Hierarchical clustering methods:

3.1. Non-Hierarchical clustering (also called k-means clustering). It first determine a cluster center, then group all objects that are within a certain distance. Examples are:

3.1.1. Sequential Threshold method - first determine a cluster center, then group all objects that are within a predetermined threshold from the center - one cluster is created at a time

3.1.2. Parallel Threshold method - simultaneously several cluster centers are determined, then objects that are within a predetermined threshold from the centre's are grouped

3.1.3. Optimizing Partitioning method - first a non-hierarchical procedure is run, then objects are reassigned so as to optimize an overall criterion.

3.2. Hierarchical clustering objects are organized into a hierarchical Structure as part of the procedure. Examples are:
3.2.1. Divisive clustering - start by treating all objects as if they are part of a single large cluster, then divide the cluster into smaller and smaller clusters.
3.2.2. Agglomerative clustering - start by treating each object as a separate cluster and then group them into bigger and bigger clusters. Examples are:
Centroid methods - clusters are generated that maximize the distance between the centers of clusters (a centroid is the mean value for all the objects in the cluster)
Variance methods - clusters are generated that minimize the within-cluster variance. Example is:
Ward's Procedure - clusters are generated that minimize the squared Euclidean distance to the center mean
Linkage methods - cluster objects based on the distance between them. Examples are:
Single Linkage method - cluster objects based on the minimum distance between them (also called the nearest neighbor rule).
Complete Linkage method - cluster objects based on the maximum distance between them (also called the furthest neighbor rule)
Average Linkage method - cluster objects based on the average distance between all pairs of objects (one member of the pair must be from a different cluster)
In order to increase similarity, objects of similar properties are placed in one class and a single access to the disk makes the entire class available. For example, in a library books concerning a large variety of topics are available. They are always kept in form of clusters. The books that have some kind of similarities among them are placed in the same cluster.. To reduce the complexity shelves are labeled with names. So when a user wants a book of specific kind on specific topic, he or she would only have to go to that particular shelf and check for the book rather than checking the entire library. Different methods are used for clustering. First is Partitioning, in which classes are mutually exclusive. Second is clumping in this case overlap is allowed. Each object is a member of cluster with which it is most similar and third is hierarchical where it produces a set of nested clusters in which each pair of objects or clusters is progressively nested in larger cluster until only one cluster remains.One example of clustering discussed here is similarity searching in medical database. This is the major application of clustering technique. In order to detect many diseases like tumour etc, the scanned pictures or x-rays are compared with the existing ones and the dissimilarities are recognized. Here clusters of images of different parts of body are stored. For example the images of the CT scan of brain are kept in cluster. To further arrange things, the image in which the right side of the brain is damaged is kept in one cluster. Hierarchical clustering is used. The stored images have already been analyzed and a record is associated with each image. When query image comes, it is firstly recognized with particular cluster image and then by similarity matching with healthy image of that specific cluster the main damaged portion or the diseased portion is recognized. Then image is sent to a specific cluster and matched with all images in that particular image with which query image has most similarities is retrieved and record associated with it is associated with query image. Thus the disease is detected.

IV. Association Rules

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics. As opposed to Sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.
Some well known algorithms of association rules are Apriori, Eclat and FP-Growth, but they only do half the job, since they are algorithms for mining frequent itemsets. Another step needs to be done after to generate rules from frequent item sets found in a database.
4.1.Apriori algorithm Apriori is the best-known algorithm to mine association rules. It uses a breadth-first search strategy which uses divide- and-conquer solution reconstruction that reduces memory requirements [11].
4.2.Eclat algorithm is a depth-first search algorithm using set intersection.
4.3.FP-growth algorithm FP-growth (frequent pattern growth) uses an extended prefix-tree (FP-tree) structure to store the database in a compressed form. FP-growth adopts a divide-and-conquer approach to decompose both the mining tasks and the databases. It uses a pattern fragment growth method to avoid the costly process of candidate

generation and testing used by Apriori**.**

4.4.GUHA is a general method for exploratory data analysis that has theoretical foundations. GUHA method mines for generalized association rules using fast bitstrings operations. The association rules mined by this method are more general than those output by apriori, for example "items" can be connected both with conjunction and disjunctions and the relation between antecedent and consequent of the rule is not restricted to setting minimum support and confidence as in apriori an arbitrary combination of supported interest measures can be used.

4.5.OPUS search is an efficient algorithm for rule discovery that, in contrast
to most alternatives, does not require either monotone or anti-monotone constraints such as minimum support Initially used to find rules for a fixed consequent, it has subsequently been extended to find rules with any item as a consequent OPUS search is the core technology in the popular Magnum Opus association discovery system.

4.6.Lore A famous story about association rule mining is the "beer and diaper" story. A purported survey of behaviour of supermarket shoppers discovered that customers (presumably young men) who buy diapers tend also to buy beer. This story became popular as an example of how unexpected association taxonomy reflects the emerging role of data visualization as a separate data. Different data mining tasks are grouped into categories depending on the type of knowledge extracted by the tasks. Profiles by improving customization and provides user with pages, and advertisements of interest. It also involves internet advertisement and fraud detection. The Web has established itself as the largest public data repository ever available.

V. Data Mining Tasks

Data mining is the process of searching and analyzing data in order to find implicit, but potentially useful, information. It involves selecting, exploring and modelling large amounts of data to uncover previously unknown patterns, and ultimately comprehensible information, from large databases.

Pattern extraction is an important component of any data mining activity and it deals with relationships between subsets of data.

Data mining tasks are used to extract patterns from large data sets. The various data mining tasks can be broadly divided into five categories as summarized in Fig.3. The taxonomy reflects the emerging role of data visualization as a separate data. Different data mining tasks are grouped into categories depending on the type of knowledge extracted by the tasks. Profiles by improving customization and provides user with pages, and advertisements of interest. It also involves internet advertisement and fraud detection. The Web has established itself as the largest public data repository ever available.

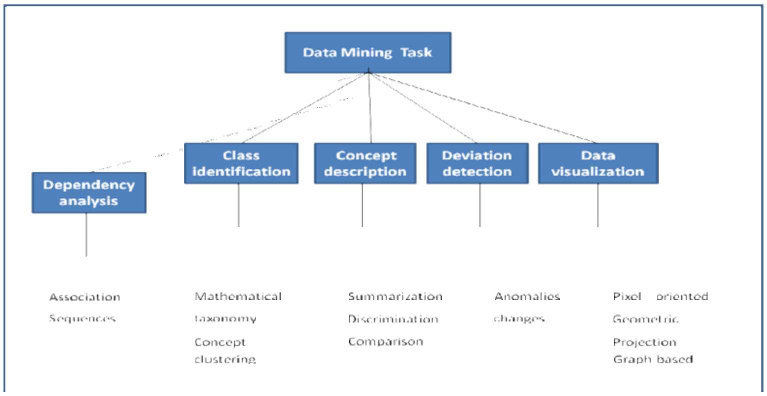


Fig 3: Data Mining Tasks

5.1. Dependency analysis

The primary type of dependency knowledge is the association between sets of items stated with some minimum specified confidence. This is also called a “market basket analysis” and gives us the relationship between different products purchased by a customer.

5.2. Class identification

Class identification groups customers into classes, which are defined in advance. There are two types of class identification tasks — mathematical taxonomy and concept clustering. Mathematical taxonomy algorithms produce classes that maximize similarity within classes but minimize similarity between classes. Concept clustering determines clusters according to attribute similarity as well as conceptual cohesiveness as defined by domain knowledge. Users provide the domain knowledge by identifying useful clustering characteristics.

5.3. Concept description

Concept description is a technique to group customers based on domain knowledge and the database, without forced definitions of the groups. Concept description can be used for summarization, discrimination, or comparison. Comparison analysis can be done by statistical or visualization technique

5.4. Deviation detection

Deviations are useful for the discovery of anomaly and changes. Anomalies are things that are different from the normal. Anomalies can be detected by analysis of the means, standard deviations, and volatility measures from the data. Confirmation of the A change B is made after investigation and the knowledge is updated.

5.5. Data visualization

To explore the knowledge in database, data visualization can be used alone or in association with other tasks such as dependency analysis, class identification, concept description and deviation detection.

VI. Data Mining Tools

The development and application of data mining algorithms requires the use of powerful software tools. Many advanced tools for data mining are available either as open-source or commercial software. Different criteria's are used to classify data mining tools based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. The 10 most frequently used tools for 'Data Mining' were RapidMiner, R, Excel, KNIME, WEKA, SAS, MATLAB, IBM SPSS Statistics, IBM SPSS Modeler, and Microsoft SQL Server [7].

WEKA WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It is freely available under the GNU General public License. It works on almost all operating systems. WEKA class libraries can be run on any computer with a web browsing capability. WEKA can also be used as a stand-alone learner. A user can apply machine learning techniques to their own data. Tools are provided for pre processing data, feeding it to a variety of learning schemes, and analyzing the resulting classifiers and their performance. It is collection of data preprocessing and modeling techniques. WEKA provides access to SQL databases using Java databse connectivity and can process the result returned by database query. WEKA includes algorithms for learning association rules and clustering. WEKA has programs that user can use to preprocess their data in order to improve learning performance or to put data into the format numeric or nominal required by particular learning algorithms.

The menu consists of five sections: Program, Applications, Tools, Visualization and Help.

1. Program Program menu having two submenus LogWindow and Exit.

Program is used to opens a log window that captures all that is printed to stdout or stderr. Useful for environments like MS Windows, where WEKA is normally not started from a terminal. Exit Closes WEKA.

2. Applications Lists the main applications within WEKA. It have 4 submenus Explorer An environment for exploring data with. Experimenter An environment for performing experiments and conducting statistical tests between learning schemes. KnowledgeFlow This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning. SimpleCLI Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

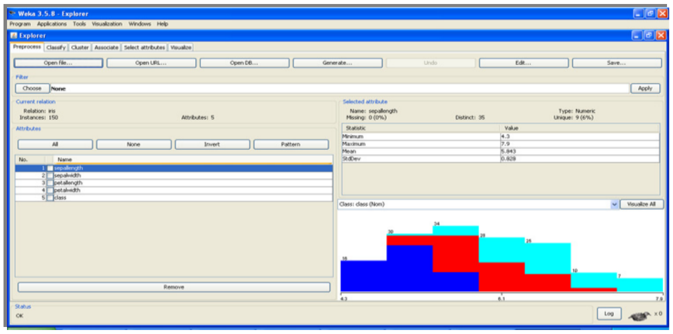
3. Tools Other useful applications are ArffViewer An MDI application for viewing ARFF files in spreadsheet format. SqlViewer represents an SQL worksheet, for querying databases via JDBC.

4. Visualization Ways of visualizing data with WEKA. Plot For plotting a 2D plot of a dataset. ROC Displays a previously saved ROC curve. TreeVisualizer For displaying directed graphs, e.g., a decision tree. GraphVisualizer Visualizes XML BIF or DOT format graphs, e.g., for Bayesian networks. BoundaryVisualizer Allows the visualization of classifier decision boundaries in two dimensions. Windows All open windows are listed here. Minimize Minimizes all current windows. Restore Restores all minimized windows again.

5. Help Online resources for WEKA can be found here. Weka homepage Opens a browser window with WEKA's home-page. Online documentation Directs to the WekaDoc Wiki [4]. HOWTOs, code snippets, etc. The general WekaWiki [3], con-taining lots of examples and HOWTOs around the development and use of WEKA. Weka on SourceforgeWEKA's project homepage on Sourceforge.net. SystemInfo Lists some internals about the Java/WEKA environ-ment, e.g., the CLASSPATH. About The infamous "About"

The WEKA Explorer

At the very top of the window, just below the title bar, is a row of tabs. When the Explorer is first started only the first tab is active; the others are greyed out. This is because it is necessary to open (and potentially pre- process) a data set before starting to explore the data. The tabs are as follows:


Weka 3.5.8 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose
Apply
Current relation
Attributes
All None Invert Pattern
Remove
Selected attribute
Visualize All
30
34
28
10
7
4.3
6.1
7.9
Status
Log
x 0

Filter: None

Relation: iris
Instances: 150
Attributes: 5

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

OK

Fig. 4. The WEKA Explorer

1. Pre-process. Choose and modify the data being acted on.
2. Classify. Train and test learning schemes that classify or perform regression.
3. Cluster. Learn clusters for the data.
4. Associate. Learn association rules for the data.
5. Select attributes. Select the most relevant attributes in the data.
6. Visualize. View an interactive 2D plot of the data.

Selecting a Classifier

At the top of the classify section is the Classifier box. This box has a text field that gives the name of the currently selected classifier, and its options.

The Choose button allows to choose one of the classifiers that are available in WEKA. Listed as: Naive Bayes, ADTree, J48, Functions, Lazy, Meta, Mi, Misc, Trees, Rule. The **Naive Bayes** Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. It is Simple and more sophisticated classification methods. Naive Bayes classifier assumes the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Example : a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. An advantage of the naive Bayes classifier is it only requires a small amount of training data to estimate the parameters necessary for classification. A Bayesian network (BN) consists of a directed acyclic graph G and a set P of probability distributions. Nodes and arcs in G represent random variables and direct correlations between variables respectively. P is the set of local distributions for each node. A local distribution is typically specied by a conditional probability table (CPT). **ADTree** (An alternating decision tree) is a machine learning method for classification which generalizes decision trees. An alternating decision tree consists of two nodes. Decision nodes specify a predicate condition and Prediction nodes contain a single number. ADTree always have prediction nodes as both root and leaves. A precondition Evaluation of a rule involves a pair of nested if statements: **Simple Cart:** Simple Cart is a classification technique that generates the binary decision tree. Simple Cart handles the missing data by ignoring that record. This algorithm is best for the training data. **J48** decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset. **ZeroR** is classifier predicts the majority of class in training data. It predicts the mean for numeric value & mode for nominal class. Random forest is a classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. For many data sets, it produces a highly accurate classifier. This algorithm handles the missing data & maintains the accuracy Decision Table. This is Class for building and using a simple decision table majority classifier In this algorithm, we have to set the number folds, display rule to get the proper result. **Random forest** is a classifier algorithm. This algorithm handles the missing data & maintains the accuracy Decision Table: This is Class for building and using a simple decision table majority classifier. In this algorithm, we have to set the number folds, display rule to get the proper result.

Selecting a Cluster

COBWEB : It is an incremental system for hierarchical conceptual clustering. COBWEB incrementally organizes observations into a classification tree. Each node in a classification tree represents a class and is labelled by a probabilistic concept that summarizes the attribute-value distributions of objects classified under the node. Classification tree can be used to predict missing attributes or the class of a new object. COBWEB performs four basic operations : Merging Two Nodes, Splitting a node , Inserting a new node. Passing an object down the hierarchy.**DBSCAN** (Density-Based Spatial Clustering Of Applications With Noise) DBSCAN is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN's definition of a cluster is based on the idea of density reachability. A point is directly density-reachable from a point if it is not farther away than a given distance .A point P is directly density-reachable from a point Q if it is not farther away than a given distance. **K-Means** is a well known iterative distance-based clustering algorithm; it also is one of the oldest, simplest and most widely used clustering algorithms. **EM** is a statistical model that makes use of the finite Gaussian mixture. **Farthest- First** combines hierarchical clustering and distance-based clustering. It also uses the basic concept of agglomerative hierarchical clustering in combination with a distance measurement criterion that is similar to the one used by K-Means.

Selecting a Associate

Apriori Association rule Used to mine the frequent patterns in database.
Support & confidence are the normal method used to measure the quality of association rule. Support is the percentage of transaction in the database that contains XUY. Confidence is the ratio of the number of transaction that
contains XUY to the number of transaction that contain X. Purpose of this
algorithm is to find subsets which are common to at least a minimum number C(Confidence Threshold) of the item sets.

Predictive Association rule In this rule support & confidence is combined
into a single measure called predictive accuracy. This predictive accuracy is used to generate the Apriori association rule. In WEKA, this algorithm generates „n‟ best association rule based on n selected by the user.

Tertius Association Rule This algorithm finds the rule according to the
confirmation measures . It uses first order logic representation. Includes various option like class Index, classification, confirmation Threshold, confirmation Values, frequency Threshold, horn Clauses, missing Values, negation, noise Threshold, number Literals, repeat Literals, roc Analysis, values Output etc.

Filtered Association rule In this algorithm data is passed through an arbitrary filter. The structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure. Here in this algorithm we can consider the Apriori, Predictive Apriori & Tertius association rule algorithm to get the result

HotSpot Algorithm This is an association rule mining algorithm which is directed by a target attribute. Which means that the RHS or consequent is fixed to the target attribute. It can be used for segmentation with both nominal and numeric targets, where the LHS would define the segment characteristics for segments which are significantly different from the entire dataset in terms of the target attribute. It uses a greedy approach to construct the tree of rules in a depth-first fashion.

ORANGE Orange is open source with active community. We can freely browse and access the source code, extend and reuse it, participate in its development, the community provides with the support, guidance and ideas. Orange runs on Windows, Mac OS X, and variety of Linux operating systems. Figure 5 Shows GUI of Orange.

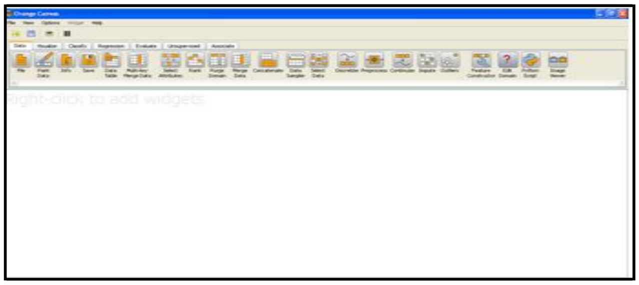

Orange Canvas
File View Options Widget Help
Data Visualize Classify Regression Evaluate Unsupervised Associate
Right-click to add widgets

Fig. 5. ORANGE GUI

Features of Orange are Visual Programming, Visualization, Interaction and Data Analytics, Large Toolbox, Scripting Interface, Extendable, Documentation, Open Source, Platform Independence and many more. We can design data analysis process through visual programming. Orange remembers choices, suggests most frequently used combinations, and intelligently chooses which communication channels between widgets to use. Orange widgets are Data, Visualize, Classify, Regression, Evaluate, Associate, Unsupervised. Various widgets are shown in figure 6 given bellow and using these widgets we can do an application like classification application shown in figure7. Orange is packed with different visualizations, from scatter plots, bar charts, trees, dendrograms, networks and heat maps. Dendrogram of an application is shown in Figure 8 bellow. Also specialized add-ons are available, like Bio orange for bioinformatics. With scripting interface (Figure 9) in Python, programming new algorithms and developing complex data analysis procedures is easy.

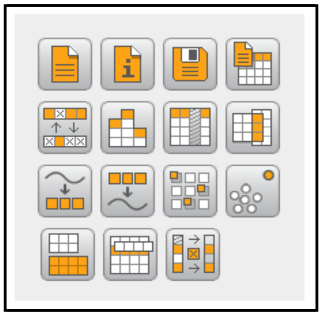


Fig 6. Orange widgets

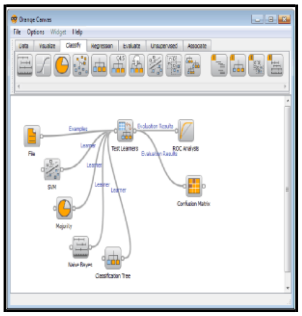


Fig 7. Classification

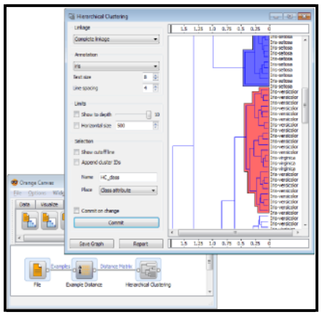
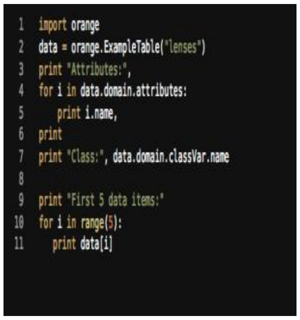


Fig 8. Dendrogram



```
import orange
data = orange.ExampleTable("lenses")
print "Attributes:",
for i in data.domain.attributes:
    print i.name,
print
print "Class:", data.domain.classVar.name

print "First 5 data items:"
for i in range(5):
    print data[i]
```

Fig 9. Scripting Interface

VII. Applications of Data Mining

7.1. Spatial Mining

The explosive growth of spatial data and widespread use of spatial databases emphasize the need for the automated discovery of spatial knowledge. Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from spatial databases [12]. The complexity of spatial data and intrinsic spatial relationships limits the usefulness of conventional data mining techniques for extracting spatial patterns. Basic tasks of spatial data mining are Classification, Association rules, Characteristic rules, Discriminate rules, Clustering, Trend detection [13]. Figure 10 shows the working of spatial data mining.

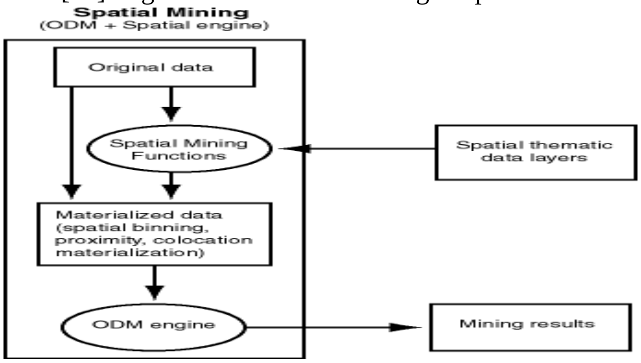

Spatial Mining
(ODM + Spatial engine)
Original data
Spatial Mining Functions
Spatial thematic data layers
Materialized data (spatial binning, proximity, colocation materialization)
ODM engine
Mining results

Fig. 10. Spatial mining and Oracle Data Mining

The original data, which included spatial and non spatial data, is processed to produce materialized data. Spatial data in the original data is processed by spatial mining functions to produce materialized data. The processing includes such operations as spatial binning, proximity, and collocation materialization. The Oracle data mining (ODM) engine processes materialized data (spatial and non spatial) to generate mining results.

7.2. Multimedia Data Mining

The exploration and analysis by automatic or semi-automatic means of large quantities of audio, video, image and text data together in order to discover meaningful patterns and rules. A multimedia database system stores and manages a large collection of multimedia data, such as audio, video, image, graphics, speech, text, document, and hypertext data, which contain text, text markups, and linkages Similarity Search in Multimedia Data When searching for similarities in multimedia data, we can search on either the data description or the data content approaches Colour histogram–based signature. Multifeature composed signature, Wavelet-based signature, and Wavelet-based signature with region-based granularity.

7.3. Text Mining

Text Data Analysis and Information Retrieval (IR) is a field that has been developing in parallel with database systems for many years. Basic Measures for Text Retrieval are Precision and Recall Precision: This is the percentage of retrieved documents that are in fact relevant to the query (i.e., "correct" responses).Recall: This is the percentage of documents that are relevant to the query and were, in fact, retrieved.

7.4. Web Data Mining

Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. Web data mining can be defined as the discovery and analysis of useful information from the WWW (World Wide Web) data. Web involves three types of data; data on the WWW, the web log data regarding the users who browsed the web pages and the web structure data. The Web data mining should focus on three issues; web structure mining, web content mining and web usage mining. Web structure mining involves mining the web document's structures and links. Web structure mining is very useful in generating information such visible web documents, luminous web documents and luminous paths; a path common to most of the results returned. Web content mining describes the automatic search of information resources available on- line. Web usage mining includes the data from server access logs, user registration or profiles, user sessions or transactions etc [14]. Other data mining applications are in Banking and Finance, in Retail, in Healthcare, in Telecommunications etc [15].

VIII. Conclusion

Today data mining is widely used in variety of applications. This paper explains how different data mining tasks can be applied to real data by giving applications. In this paper KDD process is explained in detail. We also explain different Data mining techniques like classification, clustering, association rule. This is followed by explaining various data mining tasks. Data Mining tools WEKA and ORANGE are also exploited in detail. Applications of data mining like spatial mining, multimedia mining, text mining and web data mining are dealt with.

References

[1] Osama K. Solieman, Data Mining in Sports: A Research Overview , MIS Masters Project ,August 2006.

[2]http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

[3] Michael J. Shaw, Knowledge management and data mining for marketing a,b,c,), Chandrasekar Subramaniam a, Gek Woo Tan a, Michael E. Welge b

[4]"Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis" Shelly Gupta, AIM & ACT, Banasthali University, Student M.Tech. (CS), Banasthali, India. Dharminder Kumar Dean, Faculty of Engineering and Technology, GJUS&T Hisar, India. Anand Sharma Department of CSE, GJUS&T, Project Fellow, Hisar, India.

[5]http://databases.about.com/od/datamining/g/classification.htm

[6] Data Clustering and Its Applications, Raza Ali (425), Usman Ghani (462), Aasim Saeed (464)

[7] Data mining tools, Ralf Mikut* and Markus Reischl.

[8]http://en.wikipedia.org/Association_rule_learning

[9] http://www.web-datamining.net/

[10]Web Mining Applications in E-Commerce and E-Services. ISBN 978-3-540-88080-6. Studies in Computational Intelligence, Vol. 172. Ting, I- Hsien; Wu, Hui-Ju (Eds.) 2009, VIII, 182 p. 54 illus.

[11] Breadth-First Heuristic Search ,14th International Conference on Automated Planning and Scheduling (ICAPS-04) Whistler, British Columbia, Canada ^{2} June 3 - 7, 2004. Rong Zhou and Eric A. Hansen. Department of Computer Science and Engineering Mississippi State University, Mississippi State, MS 39762

[12]Trends in Spatial Data Mining by, Shashi Shekhar, Pusheng Zhang, Yan Huang, Ranga Raju Vatsavai

[13] N.Sumathi, R.Geetha, Dr. S. Sathiya Bama. SPATIAL DATA MINING - TECHNIQUES TRENDS AND ITS APPLICATIONS Journal of Computer Applications, Vol – 1, No.4, Oct – Dec 2008

[14] Sanjay Madria, Sourav S Bhowmick, w. -k ng, e. P. Lim. Research Issues in Web Data Mining.

[15] Industry Applications of Data Mining available on site http://www.pearsonhighered.com/samplechapter/0130862711.pdf.
Accessed on date 7/8/12.