MARUDHAR KESARI JAIN COLLEGE FOR WOMEN

SUBJECT NAME: DATA MINING

SUBJECT CODE: CECA54A

Unit-5: DATA MININ METHODOLOGIES Teaching Hours: 7 Hrs. Other Methodologies of Data Mining – Data Mining Applications – Data Mining Trends – Recent Data Mining Tools – Rapid miner – Orange – Weka–Knime–Sisense – Ssdt (SQL Server Data Tools) – Oracle – Rattle – Data melt – Apache Mahout

Unit -v

Trends in Data Mining

Businesses that have been slow in adopting the process of data mining are now catching up with the others. Extracting important information through the process of data mining is widely used to make critical business decisions. We can expect data mining to become as ubiquitous as some of the more prevalent technologies used today in the coming decade. Data mining concepts are still evolving, and here are the following latest trends, such as:

1. Application exploration

Data mining is increasingly used to explore applications in other areas, such as financial analysis, telecommunications, biomedicine, wireless security, and science.

2. Multimedia Data Mining

This is one of the latest methods which is catching up because of the growing ability to capture useful data accurately. It involves data extraction from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. The data is converted into a numerical representation in different formats. This method can be used in clustering and classifications, performing similarity checks, and identifying associations.

3. Ubiquitous Data Mining

This method involves mining data from mobile devices to get information about individuals. Despite having several challenges in this type, such as complexity, privacy, cost, etc., this method has a lot of opportunities to be enormous in various industries, especially in studying human-computer interactions.

4. Distributed Data Mining

This type of data mining is gaining popularity as it involves mining a huge amount of information stored in different company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based on them.

5. Embedded Data Mining

Data mining features are increasingly finding their way into many enterprise software use cases, from sales forecasting in CRM SaaS platforms to cyber threat detection in intrusion detection/prevention systems. The embedding of data mining into vertical market software applications enables prediction capabilities for any number of industries and opens up new realms of possibilities for unique value creation.

6. Spatial and Geographic Data Mining

This new trending type of data mining includes extracting information from environmental, astronomical, and geographical data, including images taken from outer space. This type of data mining can reveal various aspects such as distance

and topology, which are mainly used in geographic information systems and other navigation applications.

7. Time Series and Sequence Data Mining

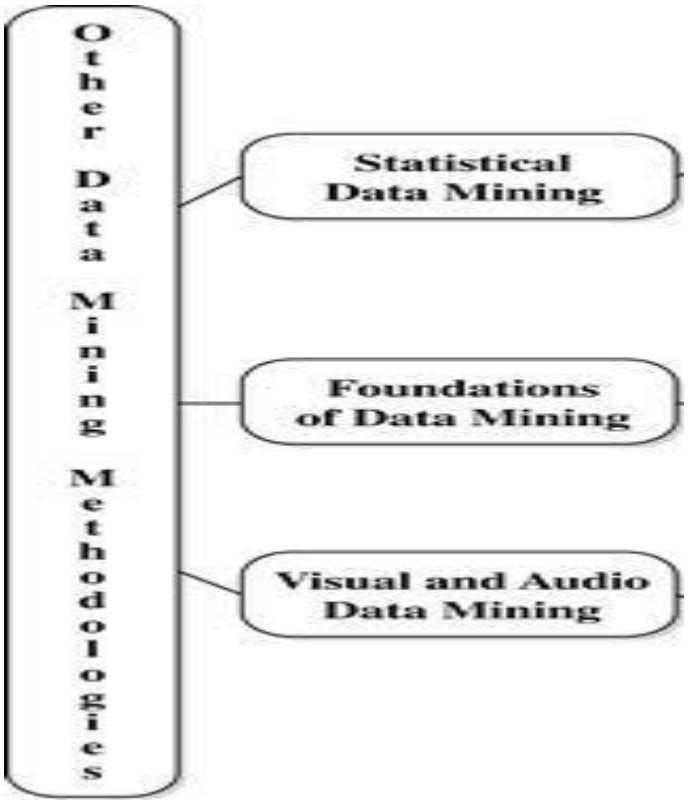
The primary application of this type of data mining is the study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events which occur outside the normal series of events. Retail companies mainly use this method to access customers' buying patterns and behaviors.

8. Data Mining Dominance in the Pharmaceutical And Health Care Industries

Both the pharmaceutical and health care industries have long been innovators in the category of data mining. The recent rapid development of coronavirus vaccines is directly attributed to advances in pharmaceutical testing data mining techniques, specifically signal detection during the clinical trial process for new drugs. In health care, specialized data mining techniques are being used to analyze DNA sequences for creating custom therapies, make better-informed diagnoses, and more.

9. Increasing Automation In Data Mining

Today's data mining solutions typically integrate ML and big data stores to provide advanced data management functionality alongside sophisticated data analysis techniques. Earlier incarnations of data mining involved manual coding by specialists with a deep background in statistics and programming. Modern



techniques are highly automated, with AI/ML replacing most of these previously manual processes for developing pattern-discovering algorithms.

10. Data Mining Vendor Consolidation

If history is any indication, significant product consolidation in the data mining space is imminent as larger database vendors acquire data mining tooling startups to augment their offerings with new features. The current fragmented market and a broad range of data mining players resemble the adjacent big data vendor landscape that continues to undergo consolidation.

11. Biological data mining

Mining DNA and protein sequences, mining high dimensional microarray data, biological pathway and network analysis, link analysis across heterogeneous biological data, and information integration of biological data by data mining are interesting topics for biological data mining research.

Other methodologies of data mining

1) Statistical data mining:

Statistical data mining techniques are created for the effective handling of large amounts of data that are generally multidimensional and possibly of several complex types.

There are various methodologies of statistical data mining are as follows −

Regression − In general, these techniques are used to forecast the value of a response (dependent) variable from new predictor (independent) variables, where the variables are numeric. There are several forms of regression, including linear, multiple, weighted, polynomial, nonparametric, and robust (robust methods are beneficial when errors declines to satisfy normalcy conditions or when the data include significant outliers).

Generalized linear models − These models and their generalization (generalized additive models), enable a categorical (nominal) response variable (several transformation of it) to be associated with a set of predictor variables in a manner same to the modeling of a mathematical response variable utilizing linear

regression. Generalized linear models involve logistic regression and Poisson regression.

Analysis of variance − These method analyze experimental information for two or more populations defined by a numeric response variable and new categorical variables (factors). In general, an ANOVA (single-factor analysis of variance) problem contains a comparison of k population or treatment defines to decide if at least two of the means are different.

Mixed-effect models − These models are for exploring grouped data—data that can be classified as per the one or more grouping variables. They generally define relationships between a response variable and several covariates in data combined according to one or more factors. There are several areas of application such as multilevel data, repeated measures data, block designs, and longitudinal data.

Factor analysis − This method can determine which variables are combined to produce a given factor. For instance, for several psychiatric data, it is not applicable to compute a specific factor of interest directly (e.g., intelligence); however, it is applicable to measure other quantities that reflect the element of interest. Therefore, none of the variables is appropriated as dependent.

Discriminant analysis − This technique can predict a categorical response variable. Unlike generalized linear models, it considers that the independent variables follow a multivariate normal distribution. The process tries to decide several discriminant functions (linear set of the independent variables) that discriminate between the groups represented by the response variable. Discriminant analysis is generally used in social sciences.

Survival analysis − There are multiple well-established statistical methods exist for survival analysis. These techniques initially were designed to forecast the probability that a patient undergoing a medical analysis can survive at least to time t.

Quality control − There are multiple statistics is used to prepare charts for quality control, including Shewhart charts and CUSUM charts. These statistics involve the mean, standard deviation, range, count, moving average, moving standard deviation, and moving range.

2) Foundations of data mining:

There are several theories for the basis of data mining include the following −

Data reduction − In this theory, the basis of data mining is to reduce the data representation. Data reduction trades certainty for speed in response to the need to obtain fast approximate answers to queries on huge databases.

Data reduction methods include singular value decomposition (the driving component behind principal components analysis), wavelets, regression, log-linear models, histograms, clustering, sampling, and the development of index trees.

Data compression − According to this theory, the basis of data mining is to compress the given information by encoding in terms of bits, association rules, decision trees, clusters, etc.

Pattern discovery − In this theory, the basis of data mining is to find patterns appearing in the database, including associations, classification models, sequential patterns, etc. There are various areas including machine learning, neural network, association mining, sequential pattern mining, clustering, and several different subfields contribute to this theory.

Probability theory − This is based on statistical theory. In this theory, the basis of data mining is to find joint probability distributions of random variables, for instance, Bayesian belief networks or hierarchical Bayesian models.

Microeconomic view – The microeconomic view considers data mining as the service of discovering patterns that are fascinating only to the extent that they can be used in the decision-making procedure of some enterprise (e.g., regarding marketing approaches and production plans).

This view is one of service, in which patterns are considered interesting if they can be based on. Enterprises are regarded as facing optimization issues, where the object is to maximize the service or value of a decision. In this theory, data mining becomes a nonlinear optimization issues.

Inductive databases − According to this theory, a database schema includes data and patterns that are saved in the database. Data mining is the problem of implementing induction on databases, where the function is to query the information and the theory (i.e., patterns) of the database. This view is famous between several researchers in database systems.

Visual and audio data mining:

Visual data mining finds implicit and beneficial knowledge from huge data sets using data and knowledge visualization methods. The human visual system is

managed by the eyes and brain, the latter of which can be think of as a dynamic, largely parallel processing and reasoning engine including a huge knowledge base.

Visual data mining can be considered as a unification of two disciplines such as data visualization and data mining. It can also associated with computer graphics, multimedia systems, human computer interaction, pattern identification, and highperformance computing.

In general, data visualization and data mining can be integrated in the following ways −

Data visualization – Data in a database or data warehouse can be considered at multiple levels of granularity or abstraction, or as several combinations of attributes or dimensions. Data can be displayed in several visual forms, including boxplots, 3-D cubes, data distribution charts, curves, surfaces, link graphs, etc.

Data mining result visualization − Visualization of data mining results is the presentation of the results or knowledge acquired from data mining in visual forms. Such forms can involve scatter plots and boxplots (acquired from descriptive data mining), and decision trees, association rules, clusters, outliers, generalized rules, etc.

Data mining process visualization – This kind of visualization presents the multiple processes of data mining in visual forms so that users can view how the data are derived and from which database or data warehouse they are extracted and how the selected data are cleaned, integrated, preprocessed, and mined. Furthermore, it can also show which approach is selected for data mining, where the results are saved, and how they can be considered.

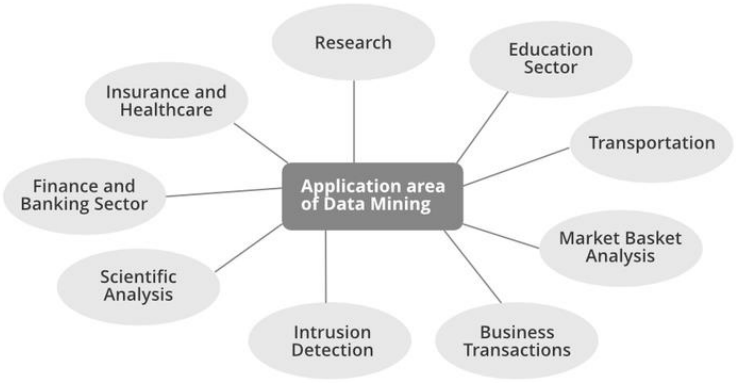
Interactive visual data mining – In interactive visual data mining, visualization tools can be used in the data mining process to provide users create intelligent data mining decisions. For instance, the data distribution in a group of attributes can be showed using colored sectors (where the whole space is defined by a circle). This display supports users decide which sector must first be selected for classification and where the best split point for this sector can be.

Audio data mining need audio signals to denote the patterns of data or the features of data mining outcomes. Although visual data mining can disclose interesting patterns utilizing graphical displays, it needs users to concentrate on watching patterns and recognizing interesting or novel characteristics inside them.

If patterns can be changed into sound and music, rather than watching pictures, it can listen to pitches, rhythms, tune, and melody to recognize anything interesting or unusual. This can relieve various burden of visual concentration and be more comfortable than visual mining. Hence, audio data mining is an interesting counterpart to visual mining.

Data mining application:

There are many measurable benefits that have been achieved in different application areas from data mining. So, let's discuss different applications of Data Mining:



Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

Sequence analysis in bioinformatics

Classification of astronomical objects

Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion

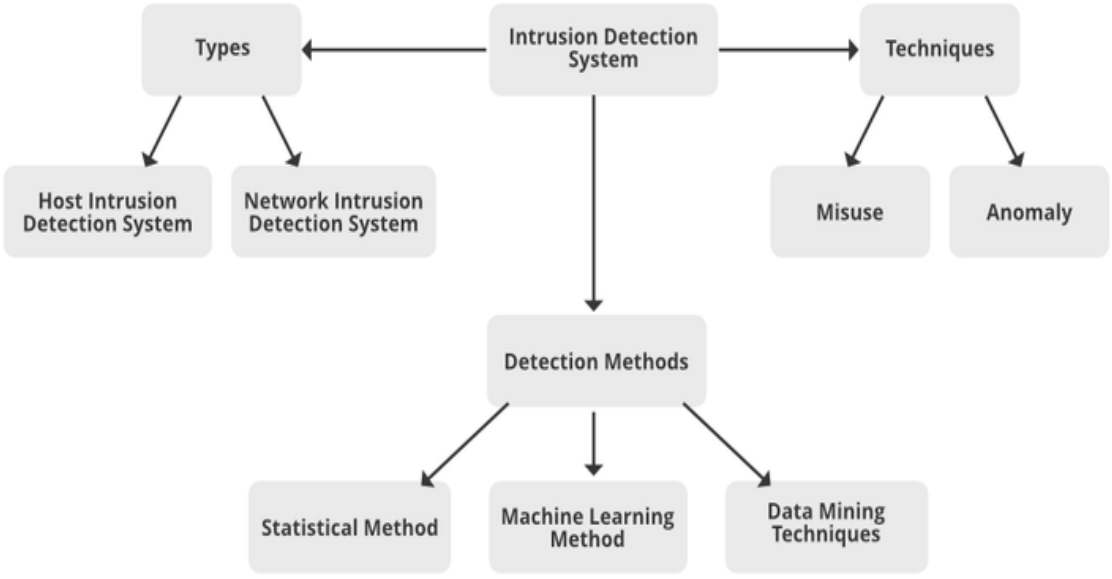
Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

Detect security violations

Misuse Detection

Anomaly Detection

Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-



business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example :

Direct mail targeting

Stock trading

Customer segmentation

Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.

Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.

Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

Predicting students admission in higher education

Predicting students profiling

Predicting student performance

Teachers teaching performance

Curriculum development

Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

Classification of uncertain data.

Information-based clustering.

Decision support system

Web Mining

Domain-driven data mining

IoT (Internet of Things)and Cybersecurity

Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

Claims analysis i.e which medical procedures are claimed together.

Identify successful medical therapies for different illnesses.

Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

Determine the distribution schedules among outlets.

Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

Credit card fraud detection.

Identify 'Loyal' customers.

Extraction of information related to customers.

Determine credit card spending by customer groups.

Data mining tools:

Data Mining tools have the objective of discovering patterns/trends/groupings among large sets of data and transforming data into more refined information.

It is a framework, such as Rstudio or Tableau that allows you to perform different types of data mining analysis.

We can perform various algorithms such as clustering or classification on your data set and visualize the results itself. It is a framework that provides us better insights for our data and the phenomenon that data represent. Such a framework is called a data mining tool.

Rapid Miner:

Rapid Miner is a data mining tool used to implement various classification and clustering algorithms. An important feature of Rapid Miner is its ability to display results visually. It is more powerful as compared to Weka because of language independance.

Rapid Miner also 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.

General fretures of Rapid Miner:

Rapid Miner is an environment for machine learning and data mining processes.

Rapid miner uses XML to describe the operator trees modelling knowledge discovery process.

It has flexible operators for data input and output file formats.

It contains more than 100 learning schemes for regression classification and clustering analysis.

Rapid Miner produces a selection of charts and visualizations automatically, choosing the most appropriate settings based on data properties

Rapid Miner supports about twenty two file formats.

Rapid Miner includes many learning algorithms in addition to WEKA.

It easily reads and writes Excel files and different databases.

If you set up an illegal work flows Rapid Miner suggest Quick Fixes to make it legal.

Rapid Miner has a responsive and intuitive GUI.

Rapid Miner is a powerful data mining tool that enables everything from data mining to model deployment, and model operations. Our end-to-end data science platform offers all of the data preparation and machine learning capabilities needed to drive real impact across your organization.

Data Importing and Exporting Tools

Rapid Miner offers dozens of different operators or ways to connect to data.

The data can be stored in a flat file such as a comma-separated values (CSV) file or spreadsheet, in a database such as a Microsoft SQLServer table, or it can be stored in other proprietary formats such as SAS or Stata or SPSS, etc.

If the data is in a database, then at least a basic understanding of databases, database connections and queries is essential in order to use the operator properly.

One could choose to simply connect to their data (which is stored in a specific location on disk) or to import the data set into the local RapidMiner repository itself so that it becomes available for any process within the repository and every time RapidMiner is opened, this data set is available for retrieval.

Either way, RapidMiner offers easy-to-follow wizards that will guide through the steps.

To simply connect to data in a CSV file on disk using a Read CSV operator, drag and drop the operator to the main process window.

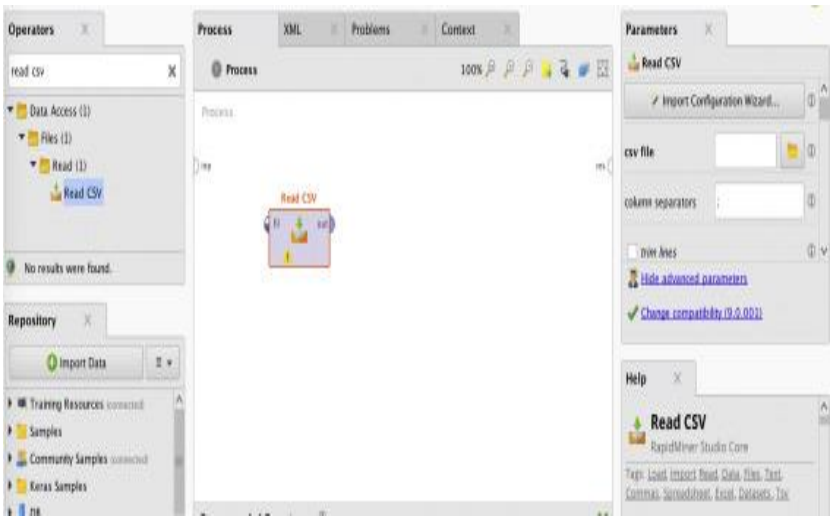
Then the Read CSV operator would need to be configured by clicking on the Import Configuration Wizard, which will provide a sequence of steps to follow to read the data in2.

The search box at the top of the operator window is also useful—if one knows even part of the operator name then it's easy to find operator.RapidMiner provides such an operator.

For example, to see if there is an operator to handle CSV files, type "CSV" in the search field and both Read and Write will show up. Clear the search by hitting the red X.

Using search is a quick way to navigate to the operators if part of their name is known. Similarly try "principal" and the operator for principal component analysis can be seen, if there is uncertainty about the correct and complete operator name or where to look initially. Also, this search shows the hierarchy of where the operators exist, which helps one learn where they are.

On the other hand, if the data is to be imported into a local RapidMiner repository,



click on the down arrow "Import Data" button in the Repositories tab and select Import CSV File. The same five-step data import wizard will immediately be presented.

Orange Data Mining

Orange is a C++ core object and routines library that incorporates a huge variety of standard and non-standard machine learning and data mining algorithms. It is an open-source data visualization, data mining, and machine learning tool. Orange is a scriptable environment for quick prototyping of the latest algorithms and testing patterns. It is a group of python-based modules that exist in the core library. It implements some functionalities for which execution time is not essential, and that is done in Python.

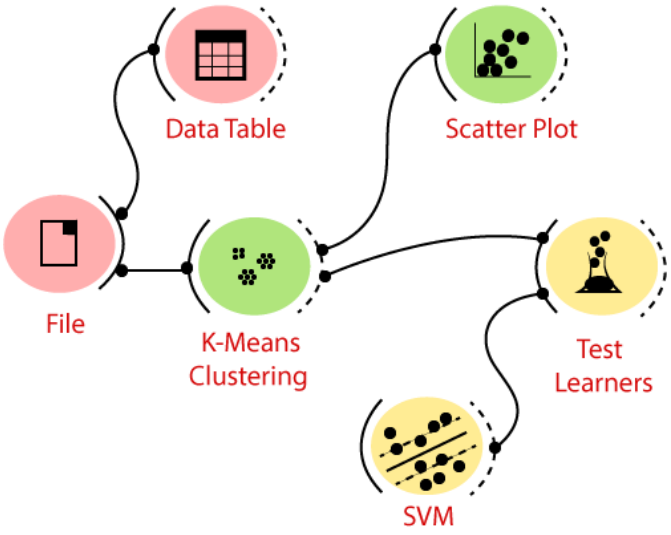
It incorporates a variety of tasks such as pretty-print of decision trees, bagging and boosting, attribute subset, and many more. Orange is a set of graphical widgets that utilizes strategies from the core library and orange modules and gives a decent user interface. The widget supports digital-based communication and can be gathered together into an application by a visual programming tool called an orange canvas.

All these together make an orange an exclusive component-based algorithm for data mining and machine learning. Orange is proposed for both experienced users and analysts in data mining and machine learning who want to create and test their own algorithms while reusing as much of the code as possible, and for those simply entering the field who can either write short python contents for data analysis.

The objective of Orange is to provide a platform for experiment-based selection, predictive modeling, and recommendation system. It primarily used in bioinformatics, genomic research, biomedicine, and teaching. In education, it is used for providing better teaching methods for data mining and machine learning to students of biology, biomedicine, and informatics.

Orange supports a flexible domain for developers, analysts, and data mining specialists. Python, a new generation scripting language and programming environment, where our data mining scripts may be easy but powerful. Orange employs a component-based approach for fast prototyping. We can implement our analysis technique simply like putting the LEGO bricks, or even utilize an existing algorithm. What are Orange components for scripting Orange widgets for visual programming?. Widgets utilize a specially designed communication mechanism for passing objects like classifiers, regressors, attribute lists, and data sets permitting to build easily rather complex data mining schemes that use modern approaches and techniques.

Orange core objects and Python modules incorporate numerous data mining tasks that are far from data preprocessing for evaluation and modeling. The operating principle of Orange is cover techniques and perspective in data mining and machine learning.



For example, Orange's top-down induction of decision tree is a technique build of numerous components of which anyone can be prototyped in python and used in place of the original one. Orange widgets are not simply graphical objects that give a graphical interface for a specific strategy in Orange, but it includes an adaptable signaling mechanism that is for communication and exchange of objects like data sets, classification models, learners, objects that store the results of the assessment. All these ideas are significant and together recognize Orange from other data mining structures.

Orange Widgets

Orange widgets give us a graphical user interface to orange's data mining and machine learning techniques. They incorporate widgets for data entry and preprocessing, classification, regression, association rules and clustering a set of widgets for model assessment and visualization of assessment results, and widgets for exporting the models into PMML.

Orange scripting:

If we want to access Orange objects, then we need to write our components and design our test schemes and machine learning applications through the script. Orange interfaces to Python, a model simple to use a scripting language with clear and powerful syntax and a broad set of additional libraries. Same as any scripting language, Python can be used to test a few ideas mutually or to develop more detailed scripts and programs.

WEKA

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modelling algorithms implemented in other programming languages, plus data preprocessing utilities in C and a makefile-based system for running machine learning experiments.

This original version was primarily designed as a tool for analyzing data from agricultural domains. Still, the more recent fully Java-based version (Weka 3), developed in 1997, is now used in many different application areas, particularly for educational purposes and research. Weka has the following advantages, such as:

Free availability under the GNU General Public License.

Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.

A comprehensive collection of data preprocessing and modelling techniques.

Ease of use due to its graphical user interfaces.

Features of Weka

Weka has the following features, such as:

1. Preprocess

The preprocessing of data is a crucial task in data mining. Because most of the data is raw, there are chances that it may contain empty or duplicate values, have garbage values, outliers, extra columns, or have a different naming convention. All these things degrade the results.

To make data cleaner, better and comprehensive, WEKA comes up with a comprehensive set of options under the filter category. Here, the tool provides both supervised and unsupervised types of operations.

2. Classify

Classification is one of the essential functions in machine learning, where we assign classes or categories to items. The classic examples of classification are: declaring a brain tumour as "malignant" or "benign" or assigning an email to a "spam" or "not_spam" class.

After selecting the desired classifier, we select test options for the training set.

3. Cluster

In clustering, a dataset is arranged in different groups/clusters based on some similarities. In this case, the items within the same cluster are identical but different from other clusters. Examples of clustering include identifying customers with similar behaviours and organizing the regions according to homogenous land use.

4. Associate

Association rules highlight all the associations and correlations between items of a dataset. In short, it is an if-then statement that depicts the probability of relationships between data items. A classic example of association refers to a connection between the sale of milk and bread.

The tool provides Apriori, FilteredAssociator, and FPGrowth algorithms for association rules mining in this category.

5. Select Attributes

Every dataset contains a lot of attributes, but several of them may not be significantly valuable. Therefore, removing the unnecessary and keeping the relevant details are very important for building a good model.

6. Visualize

In the visualize tab, different plot matrices and graphs are available to show the trends and errors identified by the model.

Knime

KNIME provides a graphical interface (a user friendly GUI) for the entire development. In KNIME, you simply have to define the workflow between the various predefined nodes provided in its repository. KNIME provides several predefined components called nodes for various tasks such as reading data, applying various ML algorithms, and visualizing data in various formats. Thus, for working with KNIME, no programming knowledge is required.

Features:

KNIME is free to use on your machine.

KNIME runs on Windows, Mac, and Linux machines.

KNIME has over 4,000 nodes for data source connections, transformations, machine learning, and visualization.

While KNIME includes a broad array of data processing and analysis capabilities, it is fully extensible by creating custom nodes using Python or R.

Many of the capabilities of H2O and WEKA also integrated and work seamlessly in the drag and drop workflow.

A variety of data file types can be used including csv, Excel, and, using an easily installed extension, databases.

Data can be exported to Excel, Tableau, Spotfire, Power BI, and other reporting platforms.

There is a large active community of users that can answer questions and provide help.

Many ready-to-use workflows are available which can be easily installed in your own work environment by dragging and dropping from the KNIME site.

Extensive documentation, learning modules, videos, and training events are available.

KNIME - Workbench

The workbench consists of several views. The views which are of immediate use to us are marked in the screenshot and listed below −

. Workspace

.Outline

. Nodes Repository

. KNIME Explorer

. Console

. Description

Workspace View

The most important view for us is the Workspace view. This is where you would create your machine learning model.

Each workspace contains one or more nodes.

The nodes are connected using arrows. Generally, the program flow is defined from left to right, though this is not required. You may freely move each node anywhere in the workspace. The connecting lines between the two would move appropriately to maintain the connection between the nodes. You may add/remove connections between nodes at any time. For each node a small description may be optionally added.

Outline View

The workspace view may not be able to show you the entire workflow at a time. That is the reason, the outline view is provided.

The outline view shows a miniature view of the entire workspace. There is a zoom window inside this view that you can slide to see the different portions of the workflow in the Workspace view.

Node Repository

This is the next important view in the workbench. The Node repository lists the various nodes available for your analytics. The entire repository is nicely categorized based on the node functions. You will find categories such as −

. IO

. Views

. Analytics

Under each category you would find several options. Just expand each category view to see what you have there. Under the IO category, you will find nodes to read your data in various file formats, such as ARFF, CSV, PMML, XLS, etc.

Depending on your input source data format, you will select the appropriate node for reading your dataset.

The Analytics node defines the various machine learning algorithms, such as Bayes, Clustering, Decision Tree, Ensemble Learning, and so on.

KNIME Explorer

The first two categories list the workspaces defined on the KNIME server. The third option LOCAL is used for storing all the workspaces that you create on your local machine. Try expanding these tabs to see the various predefined workspaces. Especially, expand EXAMPLES tab.

Console View

As the name indicates, the Console view provides a view of the various console messages while executing your workflow.

The Console view is useful in diagnosing the workflow and examining the analytics results.

Description View

The last important view that is of immediate relevance to us is the Description view. This view provides a description of a selected item in the workspace.

The view shows the description of a File Reader node. When you select the File Reader node in your workspace, you will see its description in this view. Clicking on any other node shows the description of the selected node. Thus, this view becomes

very useful in the initial stages of learning when you do not precisely know the purpose of the various nodes in the workspace and/or the nodes repository.

Sisense

Sisense is a business intelligence software company headquartered in New York City.

The company develops business intelligence software that allows users to access and analyze big data. The software uses a chip-based database engine for analytics, rather than an in-memory database engine.

Sisense is a business intelligence tool that enables organizations to “infuse analytics everywhere,” as the platform puts it, including within customer and employee applications and workflows. Sisense offers a wide range of BI tools, including data modeling, data visualization and AI analytics. The platform was designed to be easy to scale and includes security features such as attack surface monitoring and disaster recovery.

Data connections

Sisense enable businesses to integrate data sources into a single source using built-in data connectors. Sisense currently

offers over 100 data connectors, including Azure Synapse, Google BigQuery, MySQL, Snowflake and SQL Server.

It's important to mention Sisense's Elasticube. The Elasticube is Sisense's proprietary analytics cache that enables organizations to bring in data from multiple sources and manipulate it as if it was one data set. According to Sisense, the Elasticube is capable of returning queries of millions of rows of raw data in seconds.

Data visualization

Sisense offers a drag-and-drop interface for easy dashboard and visualization building. Sisense also offers interactive visualizations so companies can move beyond graphs and charts. Both platforms offer the ability to build custom dashboards and visualizations from scratch.

AI analytics

Sisense offers forecasting capabilities so users can see how a change will affect future values. The platform also offers "Simply Ask" which is a natural language query feature. Users can ask questions about their data and receive visualizations that provide the answers. Sisense returns automatic suggestions to help build queries, making this tool a great option for anyone in need of insights, regardless of skill level.

Embedded analytics

Sisense's Fusion Embed solution was designed just for building white-labeled analytic experiences inside apps. It empowers businesses to offer predictive analytics, natural language querying and more for various skill levels.

SSDT

SQL Server Data Tools (SSDT) is a modern development tool for building SQL Server relational databases, databases in Azure SQL, Analysis Services (AS) data models, Integration Services (IS) packages, and Reporting Services (RS) reports. With SSDT, you can design and deploy any SQL Server content type with the same ease as you would develop an application in Visual Studio.

Oracle

Oracle Data Mining provides a powerful, state-of-the-art data mining capability within Oracle Database. You can use Oracle Data Mining to build and deploy predictive and descriptive data mining applications, to add intelligent capabilities to existing

applications, and to generate predictive queries for data exploration.

Oracle Data Mining offers a comprehensive set of in-database algorithms for performing a variety of mining tasks, such as classification, regression, anomaly detection, feature extraction, clustering, and market basket analysis. The algorithms can work on standard case data, transactional data, star schemas, and text and other forms of unstructured data. Oracle Data Mining is uniquely suited to the mining of very large data sets.

Oracle Data Mining is one of the two components of the Oracle Advanced Analytics Option of Oracle Database Enterprise Edition. The other component is Oracle R Enterprise, which integrates R, the open-source statistical environment, with Oracle Database. Together, Oracle Data Mining and Oracle R Enterprise constitute a comprehensive advanced analytics platform for big data analytics.

Advantages:

Data mining within Oracle Database offers many advantages:

No Data Movement: Some data mining products require that the data be exported from a corporate database and converted to a specialized format for mining. With Oracle Data Mining, no data movement or conversion is needed. This makes the entire mining process less complex, time-consuming, and error-prone, and it allows for the mining of very large data sets.

Security: Your data is protected by the extensive security mechanisms of Oracle Database. Moreover, specific database privileges are needed for different data mining activities. Only users with the appropriate privileges can define, manipulate, or apply mining model objects.

Data Preparation and Administration: Most data must be cleansed, filtered, normalized, sampled, and transformed in various ways before it can be mined. Up to 80% of the effort in a data mining project is often devoted to data preparation. Oracle Data Mining can automatically manage key steps in the data preparation process. Additionally, Oracle Database provides extensive administrative tools for preparing and managing data.

Ease of Data Refresh: Mining processes within Oracle Database have ready access to refreshed data. Oracle Data Mining can easily deliver mining results based on current data, thereby maximizing its timeliness and relevance.

Oracle Database Analytics: Oracle Database offers many features for advanced analytics and business intelligence. Oracle Data Mining can easily be integrated with other analytical features of the database, such as statistical analysis and OLAP.

Domain Environment: Data mining models have to be built, tested, validated, managed, and deployed in their appropriate application domain environments. Data mining results may need to be post-processed as part of domain specific computations (for example, calculating estimated risks and response probabilities) and then stored into permanent repositories or data warehouses. With Oracle Data Mining, the pre- and post-mining activities can all be accomplished within the same environment.

Application Programming Interfaces: The PL/SQL API and SQL language operators provide direct access to Oracle Data Mining functionality in Oracle Database.

Interfaces to Oracle Data Mining

The programmatic interfaces to Oracle Data Mining are PL/SQL for building and maintaining models and a family of SQL functions for scoring. Oracle Data Mining also supports a graphical user interface, which is implemented as an extension to Oracle SQL Developer.

Oracle Predictive Analytics, a set of simplified data mining routines, is built on top of Oracle Data Mining and is implemented as a PL/SQL package.

PL/SQL API

The Oracle Data Mining PL/SQL API is implemented in the DBMS_DATA_MINING PL/SQL package, which contains routines for building, testing, and maintaining data mining models. A batch apply operation is also included in this package.

SQL Functions

The Data Mining SQL functions perform prediction, clustering, and feature extraction. The functions score data by applying a mining model object or by executing an analytic clause that performs dynamic scoring.

Oracle Data Miner

Oracle Data Miner is an extension to Oracle SQL Developer that enables data scientists and business and data analysts to view data, rapidly build multiple machine learning models, compare and evaluate multiple models, apply them to new data, and accelerate model deployment.

Oracle Data Miner enables data scientists, “citizen data scientists,” and business and data analysts to work directly with data inside the database using a graphical “drag and drop” workflow editor. Oracle Data Miner (ODMr), an extension to Oracle SQL Developer, captures and documents in graphical analytical workflows the steps users take while exploring data and developing machine learning methodologies. ODMr workflows are useful for re-executing analytical methodologies and for sharing insights with team members. ODMr generates SQL and PL/SQL scripts and offers a workflow API for accelerating model deployment throughout the enterprise.

Rattle

Ratte is a data mining tool based on GUI. It uses the R stats programming language. Rattle exposes the statical power of R by offering significant data mining features. While rattle has a comprehensive and well-developed user interface, It has an integrated log code tab that produces duplicate code for any GUI operation.

The data set produced by Rattle can be viewed and edited. Rattle gives the other facility to review the code, use it for many purposes, and extend the code without any restriction.

Features:

File Inputs = CSV, TXT, Excel, ARFF, ODBC, R Dataset, RData File, Library Packages Datasets, Corpus, and Scripts.

Statistics = Min, Max, Quartiles, Mean, St Dev, Missing, Medium, Sum, Variance, Skewness, Kurtosis, chi square.

Statistical tests = Correlation, Wilcoxon-Smirnov, Wilcoxon Rank Sum, T-Test, F-Test, and Wilcoxon Signed Rank.

Clustering = KMeans, Clara, Hierarchical, and BiCluster.

Modeling = Decision Trees, Random Forests, ADA Boost, Support Vector Machine, Logistic Regression, and Neural Net.

Evaluation = Confusion Matrix, Risk Charts, Cost Curve, Hand, Lift, ROC, Precision, Sensitivity.

Charts = Box Plot, Histogram, Correlations, Dendrograms, Cumulative, Principal Components, Benford, Bar Plot, Dot Plot, and Mosaic.

Transformations = Rescale (Recenter, Scale 0-1, Median/MAD, Natural Log, and Matrix) - Impute (Zero/Missing, Mean, Median, Mode & Constant), Recode (Binning, Kmeans, Equal Widths, Indicator, Join Categories) - Cleanup (Delete Ignored, Delete Selected, Delete Missing, Delete Obs with Missing)

Packages:

The capabilities of R are extended through user-submitted packages, which allow specialized statistical techniques, graphical devices, as well as import/export capabilities to many external data formats. Rattle uses these packages - RGtk2, pmml, colorspace, ada, amap, arules, biclust, cba, descr, doBy, e1071, ellipse, fEcofin, fBasics, foreign, fpc, gdata, gtools, gplots, gWidgetsRGtk2, Hmisc, kernlab, latticist, Matrix, mice, network, nnet, odfWeave, party, playwith, psych, randomForest, reshape, RGtk2Extras, ROCR, RODBC, rpart, RSvgDevice, survival, timeDate, graph, RBGL, bitops.

Datamelt

DataMelt is a computation and visualization environment which offers an interactive structure for data analysis and visualization. It is primarily designed for students, engineers, and scientists. It is also known as DMelt.

DMelt is a multi-platform utility written in JAVA. It can run on any operating system which is compatible with JVM (Java Virtual Machine). It consists of Science and mathematics libraries.

Scientific libraries:

Scientific libraries are used for drawing the 2D/3D plots.

Mathematical libraries:

Mathematical libraries are used for random number generation, algorithms, curve fitting, etc.

DMelt can be used for the analysis of the large volume of data, data mining, and statistical analysis. It is extensively used in natural sciences, financial markets, and engineering.

:

Chart plotting

Statistics

Data analysis

Visualization in 2D and 3D

Data mining

Symbolic computations

Vector graphics

Neural networks

Jython, Python, Java, JRuby, Groovy

Portable application

Linear and non-linear regression

Data fit

Numeric computation

Statistical computation

Linear algebra

Symbolic computations

Symbolic regression

Text processing

Apache Mahout

A mahout is one who drives an elephant as its master. The name comes from its close association with Apache Hadoop which uses an elephant as its logo.

Hadoop is an open-source framework from Apache that allows to store and process big data in a distributed environment across clusters of computers using simple programming models.

Apache Mahout is an open source project that is primarily used for creating scalable machine learning algorithms. It implements popular machine learning techniques such as:

Recommendation

Recommendation is a popular technique that provides close recommendations based on user information such as previous purchases, clicks, and ratings.

Classification

Classification, also known as categorization, is a machine learning technique that uses known data to determine how the new data should be classified into a set of existing categories. Classification is a form of supervised learning.

Clustering

Clustering is used to form groups or clusters of similar data based on common characteristics. Clustering is a form of unsupervised learning.

Apache Mahout started as a sub-project of Apache's Lucene in 2008. In 2010, Mahout became a top level project of Apache.

Features of Mahout:

The primitive features of Apache Mahout are listed below.

The algorithms of Mahout are written on top of Hadoop, so it works well in distributed environment. Mahout uses the Apache Hadoop library to scale effectively in the cloud.

Mahout offers the coder a ready-to-use framework for doing data mining tasks on large volumes of data.

Mahout lets applications to analyze large sets of data effectively and in quick time.

Includes several MapReduce enabled clustering implementations such as k-means, fuzzy k-means, Canopy, Dirichlet, and Mean-Shift.

Supports Distributed Naive Bayes and Complementary Naive Bayes classification implementations.

Comes with distributed fitness function capabilities for evolutionary programming.

Includes matrix and vector libraries.

Apache Mahout is a highly scalable machine learning library that enables developers to use optimized algorithms. Mahout implements popular machine learning techniques such as recommendation, classification, and clustering. Therefore, it is prudent to have a brief section on machine learning before we move further.