Artificial Intelligence and Pattern Recognition Using Data Mining Algorithms

Abdulkawi Yahya Radman Al-Shamiri [1, 2],

[1] Department of Computer Science, School of Computer Science and Engineering at Hodeidah University, Al-Hodeidah, Yemen.
[2] Department of Computer Application Technology, School of Computer Science and Information Engineering at Hefei University of Technology (HFUT), Hefei, China.

Abstract
In recent years, with the existence of huge amounts of data stored in huge databases, the need for developing accurate tools for analyzing data and extracting information and knowledge from the huge and multi-source databases have been increased. Hence, new and modern techniques have emerged that will contribute to the development of all other sciences. Knowledge discovery techniques are among these technologies, one popular technique of knowledge discovery techniques is data mining which aims to knowledge discovery from huge amounts of data. Such modern technologies of knowledge discovery will contribute to the development of all other fields. Data mining is important, interesting technique, and has many different and varied algorithms; Therefore, this paper aims to present overview of data mining, and clarify the most important of those algorithms and their uses.

Key words:
Data Mining, Classification Algorithms, Regression Algorithms, Association Rules Algorithms, Clustering Algorithms, Steps of Data Mining Process.

1. Introduction

Data mining technique has become more popular in many areas of life. Where it is used to recognize patterns, these patterns used in all areas in decision making and to identify relationships. For example, A large company that has many branches, products and customers, the owner of that company wants to know, for example, what products are bestselling, or what products are bought together etc.; To find out what the company owner wants, data mining algorithms are used to recognize patterns that will help the company owner know the relationships between products and make appropriate decisions.

Data mining expands every day, so there was a need for learning this field. Where data mining algorithms are used in developing the medical field to serve humanity progress in medical aspect, such as discovering the threats and challenges faced by societies in the health field. Also, are used in analysis the text data stream to text crawling, document organization and topic detection, and news group filtering etc. Data mining is used for text mining, retail stores, financial analysis, biological data analysis, to find intrusion detection, fraud detection, and other scientific applications.

The data at the beginning are incomprehensible words and symbols, then this data goes through several operations until this data becomes a pattern for discovering knowledge. Where Figure (1) illustrates the data hierarchy. Data is incomprehensible words and symbols. If data is put in a context, it will become information. If information is put in a rule, it will become a knowledge. If a set of practical experiments is done on knowledge, it will become a wisdom.

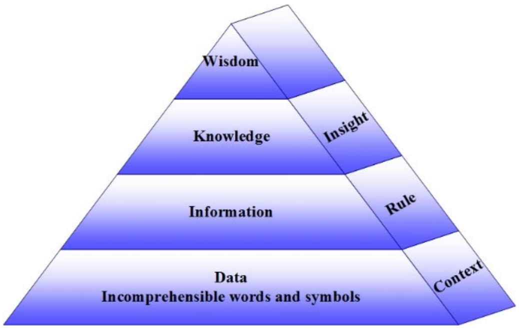


Fig. 1 Data Hierarchy.

Data mining is associated with some other computer technology, where Figure (2) illustrates the relationship of data mining with Pattern Recognition, Artificial Intelligence (AI), Machine Learning, Databases, Mathematical Modeling, Statistics, and Management Science & Information Systems.

Manuscript received July 5, 2021
Manuscript revised July 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.7.26

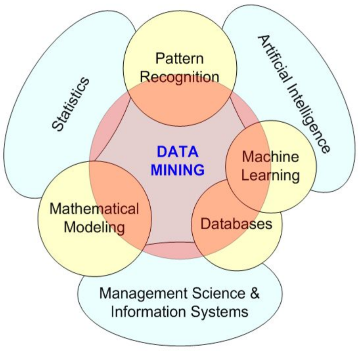


Fig. 2 Data mining at the intersection of many disciplines.

Data mining has several steps and processes. These steps begin with data collection and end with evaluation of pattern. Where Figure (3) illustrates the hierarchy of data mining process. Data mining process begins with collection of data sources, then merging data sources in a unified database. After that, the data cleaning process is done on the unified database, then outputs of the data cleaning process are put in a data warehouse. After that, choosing the appropriate data mining algorithm, then a pattern is got. Next step is the evaluation of this pattern, then knowledge is discovered.

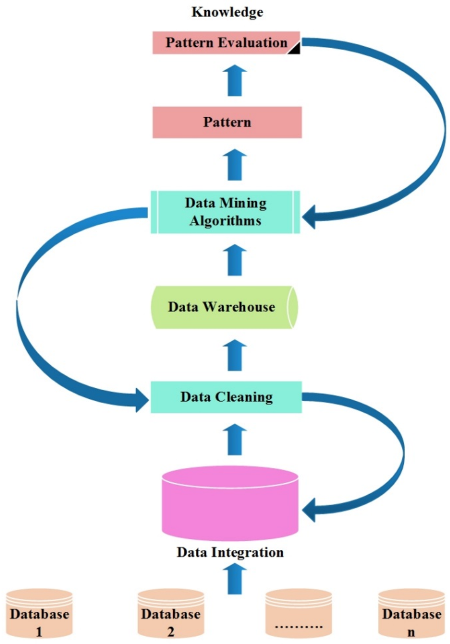


Fig. 3 Hierarchy of Data Mining Process.

This paper aims to recognize patterns by Clarification for the most important of data mining algorithms and their uses.

2. DATA MINING

2.1 What is Data Mining

Data mining is an operation of discovering knowledge from large data sets by algorithms which find out patterns of knowledge discovery. Data mining means deep study of data which are collected from several sources and to be in various formats to find out a pattern, then discover knowledge and to discover the common relationships between this data to solve a problem by this pattern. [1, 2, 3].

2.2 Difference Between KDD and Data mining

When data mining is mentioned, Knowledge Discovery in Databases (KDD) is mentioned, what is the relationship between KDD and data mining?

Knowledge Discovery in Databases (KDD), it is an area in computer science, that has theories and tools to help humanity in finding out useful and previously unknown information from large data sets, that information called knowledge. Data Mining is application of some particular algorithms to recognize patterns from large data sets. Thus, KDD is the overall process for finding out knowledge from large data sets, while Data Mining is a step of the KDD process, which deals with recognizing patterns in large data sets [4].

2.3 The data hierarchy

Figure (1) shows the data hierarchy that is as following:-

1) Data: Data is incomprehensible words and symbols. For example, (Mohamed, Ahmed, football, play).

2) Information: It is data set placed in a specific context, creating an understandable sentence. For example, we have these data (Mohamed, Ahmed, football, play), if it is putted in a context, will be Understandable information as (Mohamed and Ahmed play football together).

3) Knowledge: It is information placed in a specific rule, will create a knowledge. For example, we have this information (Mohamed and Ahmed play football together), if it is putted in a rule, will be a knowledge as (If Mohamed play football, Ahmed will play).

4) Wisdom: It is set of knowledge putted in several practical experiences, will be a wisdom.

2.4 Data mining process

Data mining process is shown in Figure (3) as a sequence of the following steps:-

1) Collection of sources: It means the collection of data sources, which are multiple databases and different in the structure and from different places.

2) Data Integration: It means Integration of the data collected from the previous step (Collection of sources) in a unified structure and one place.

3) Data Cleaning: After the process of data integration from multiple sources and different structure, will appear the following problems which need to process of data cleaning:-

a) Inconsistent Data:

It means that some data have been expressed in different methods. For example, some sources express the marital status with letters such as "S" to denote "single", and some sources express the word instead of letters, here the problem of inconsistency appears in the data. To solve this problem, you have to choose a unified format [1, 2].

b) Missing Values:

It means that you during data preparation will encounter that some data is missing. To solve this problem, there are some algorithms. You have to choose a method of following [1, 2]:-

- Delete the record that contains the missing value.
- Compensation with a random value within the range of values in other records.
- Compensation with a specified value that is found by a mathematical operation performed on other record values.

c) Noisy Data:

It is data have an amount of additional incomprehensible values. To solve this problem, there are some algorithms [1,2].

4) Data Warehouse: It is the dataset which created after data cleaning process.

5) Data Mining Algorithms: In this step, one of the various data mining algorithms is used on the data that is in the data warehouse to recognize the pattern, which will use to solve a particular problem. To recognize a pattern using data mining algorithms, one of the two methods is used (Using of one of data mining tools, or using of one of programming languages).

6) Getting the Pattern: After implementing the specified algorithm, a pattern will be generated that discovers the knowledge.

7) Pattern Evaluation: At this stage, the process of evaluating pattern that is created in the stage of implementation of the appropriate data mining algorithm is performed. The process of evaluating pattern takes place to ensure availability of the characteristics of a good pattern, where these characteristics are (Valid, Novel, Potentially useful, and Ultimately understandable) [1, 2].

2.5 Conditions for data mining

1) Existence a specialist:-

Since data mining technique is used in all fields, presence a specialist during preparing data is very important. For example, the need for a specialist in ophthalmology to prepare eye patients data to implement one of the data mining algorithms on. So, existence a specialist during preparing data is very important; to understand the data and the problems to be solved, and to interpret the results.

2) Determining the purpose:-

Determining the purpose will be before preparing data. It means to define the problem to be solved by using data mining algorithms. For example, there are data for eye patients and want to discover risk factors for blindness.

3) Understanding data:-

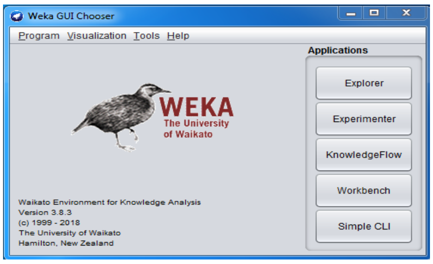
It means understanding values of data to be prepared.

3. The Most Famous Data Mining Tools

There are some data mining tools which help to Implement data mining algorithms, as the following:-

3.1 WEKA:

WEKA is an open source data mining tool, which based on Java and is compatible with all operating systems. WEKA contains various data mining algorithms such as classification, association, clustering, and regression. WEKA provides connectivity to SQL databases and can process required data from these sources. WEKA can the data mining classification with the help of decision trees and neural networks (NN). There are some weaknesses for WEKA, the cluster analysis for which only a few methods are included, and when large amounts of data have to be managed because all data is loaded into the working memory of WEKA [5].



Waikato Environment for Knowledge Analysis
Version 3.8.3
(c) 1999 - 2018
The University of Waikato
Hamilton, New Zealand

Fig. 4 The main interface of the WEKA data mining tool.

3.2 RapidMiner:

RapidMiner is available in four modules (RapidMiner Studio, RapidMiner Auto Model, RapidMiner Server and RapidMiner Radoop) as a free and a paid data mining tool, which based on Java. RapidMiner contains various data mining algorithms such as classification, association, clustering, and regression, and includes options for data, web and text mining and for sentiment analysis (sentiment analysis or opinion mining). RapidMiner can integrate Excel tables or SPSS files and data records from R-Studio and WEKA. There are some weaknesses for RapidMiner, it is difficult to handle large amounts of data, does not provide the option of comparing models created using different processes, and when used in practice, RapidMiner is relatively slow compared to the other tools [5].

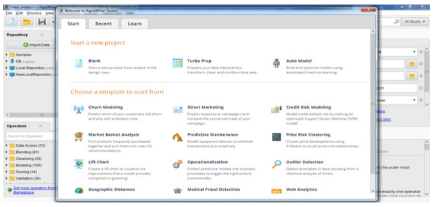


Fig. 5 The main interface of the RapidMiner data mining tool.

3.3 Orange:

Orange is an open source data mining tool, which based on C++ and is compatible with all operating systems. Access language is the Python programming language, but more complex operations are performed in C++. Orange contains various data mining algorithms such as classification, association, clustering and regression, and includes many applications for data and text analysis as well as functions for Machine Learning [5].

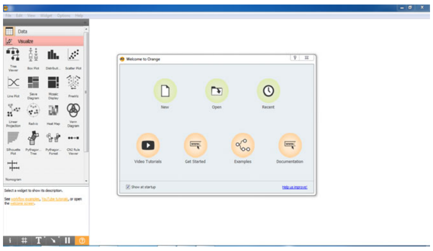


Fig. 6 The main interface of the Orange data mining tool.

3.4 KNIME:

Konstanz Information Miner (KNIME) is an open source data mining tool, which based on Java and processed with Eclipse. KNIME includes various components for machine learning and data mining through its modular data pipelining concept. KNIME's graphical user interface and use of Java Database Connectivity (JDBC) allows aggregation of nodes mixing different data sources, including preprocessing (ETL: Extraction, Transformation, Loading), for modeling, data analysis and visualization [5].

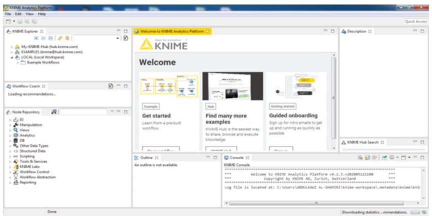


Fig. 7 The main interface of the KNIME data mining tool.

3.5 SAS:

Statistical Analysis System (SAS) is considered the leading data mining tool for business analysis, but it is the most expensive. It based on the C Programming Language. SAS' strength lies in forecasting and interactive data visualization, which can also be used for large presentations. Another strength of the data mining tool is its high scalability and performance, which can be expanded by adding hardware or other resources. Less technically savvy users benefit from the graphical user interface. However, SAS can only be used using the SAS license [5].

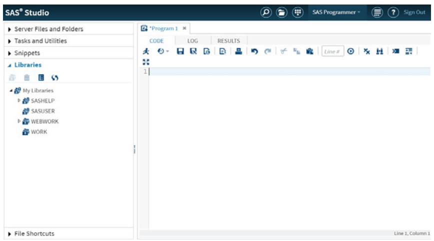


Fig. 8 The main interface of the SAS data mining tool.

3.6 Oracle Data Mining:

Oracle Data Mining (ODM) is one of Oracle Database Enterprise. ODM contains several data mining and data analysis algorithms as associations, classification, regression, prediction and specialized analytics. It provides methods for the creation, management and operational deployment of data mining models inside the database environment [6].

Fig. 9 The main interface of the Oracle data mining tool.

4. The Most Famous Programming Languages for Data Mining

You can implement data mining algorithms through programming languages by programming codes for algorithms, where this method helps you to amend and add to a specific algorithm and also enables you to build a new algorithm. There are some programming languages which help to Implement data mining algorithms, as the following:-

4.1 Python Programming Language:

Python is an easy programming language, and is wide and flexible, has many of inbuilt programming libraries, and includes a lot of algorithms, that will help in easing implement data mining algorithms. Python is a fast programming language for data mining and more practical to create a pattern [7].

4.2 Java Programming Language:

Java is an old and well-known language, and used in expansion of social media sites such as LinkedIn, Twitter and Facebook. It able to code different types of algorithms [7].

4.3 C++ Programming Language:

C++ is a high-level and general-purpose programming language. C++ provides facilities for low-level memory manipulation, and has imperative, object-oriented, and generic programming features.

4.4 R - Programming Language:

R is a free programming language, used in data analysis, data mining, graphics and statistical computing. R provides tools for data mining and analysis, and allows design high-level graphics [7].

4.5 MATLAB Programming Language:

MATLAB is a shortcut to Matrix Laboratory. It is a multi-paradigm programming language. MATLAB allows implementation of algorithms, matrix processing, creation of user interfaces, plotting of functions and data, and interfacing with programs written in other languages [8].

5. DATA MINING ALGORITHMS

There are many of different and multi-tasking data mining algorithms. Many, but not all, of these algorithms will be mentioned in this paper. The data mining algorithms will be classified as follows:-

5.1 Classification Algorithms:

1) What is classification

It is a process of creating a pattern (Classifier) that classifies data into categories (classes). For example, email messages are categorized into legitimate messages and spam messages, where this pattern determines which category the new data belongs to it, based on the previous test that was made on previously available data.

2) Mechanism of classification algorithms

Classification algorithms are based on rule (IF-THEN). A data set contains a set of records and fields (Attributes). One of these attributes is set as a class. Where this data set is divided into (training data, test data). Training data is used to train the algorithm on it, while test data are used to test the accuracy and quality of the algorithm by comparing the new results of the test data with the pre-existing values.

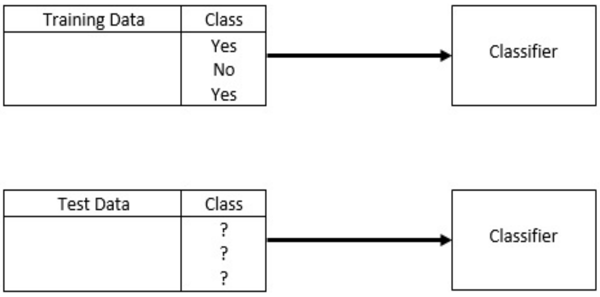
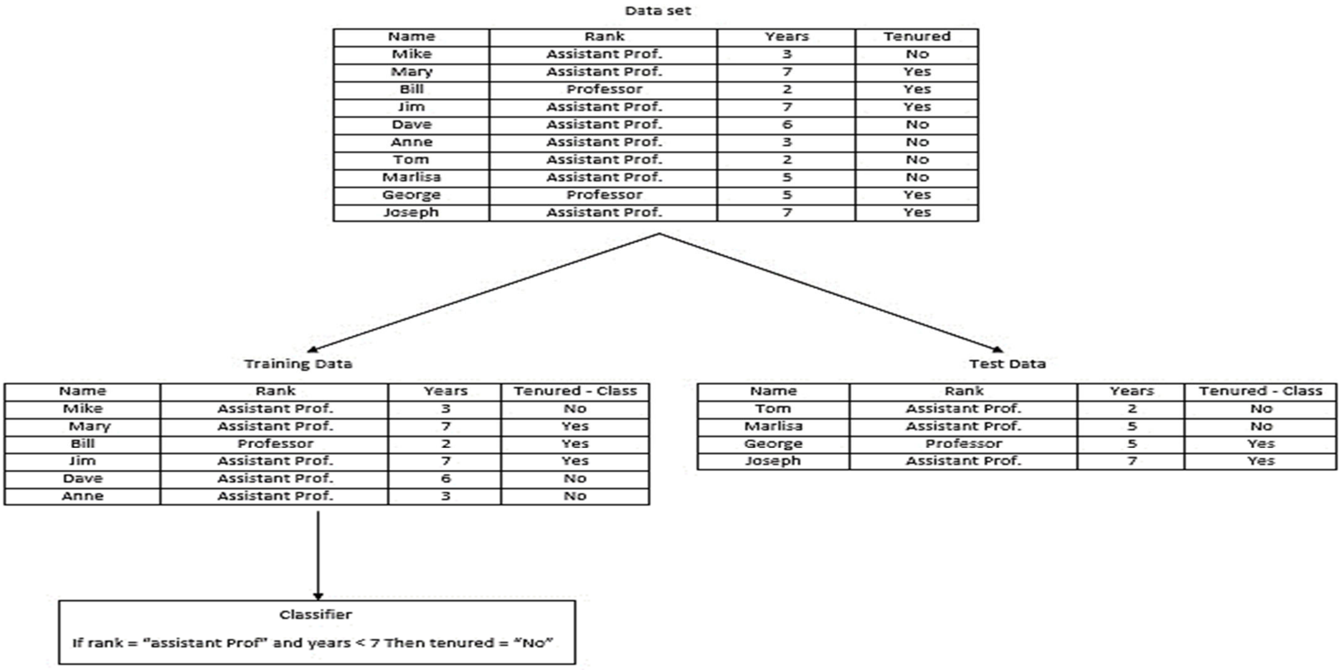


Fig. 10 Mechanism of classification algorithms.

3) Example



Data set

Name	Rank	Years	Tenured
Mike	Assistant Prof.	3	No
Mary	Assistant Prof.	7	Yes
Bill	Professor	2	Yes
Jim	Assistant Prof.	7	Yes
Dave	Assistant Prof.	6	No
Anne	Assistant Prof.	3	No
Tom	Assistant Prof.	2	No
Marlisa	Assistant Prof.	5	No
George	Professor	5	Yes
Joseph	Assistant Prof.	7	Yes

Training Data

Name	Rank	Years	Tenured - Class
Mike	Assistant Prof.	3	No
Mary	Assistant Prof.	7	Yes
Bill	Professor	2	Yes
Jim	Assistant Prof.	7	Yes
Dave	Assistant Prof.	6	No
Anne	Assistant Prof.	3	No

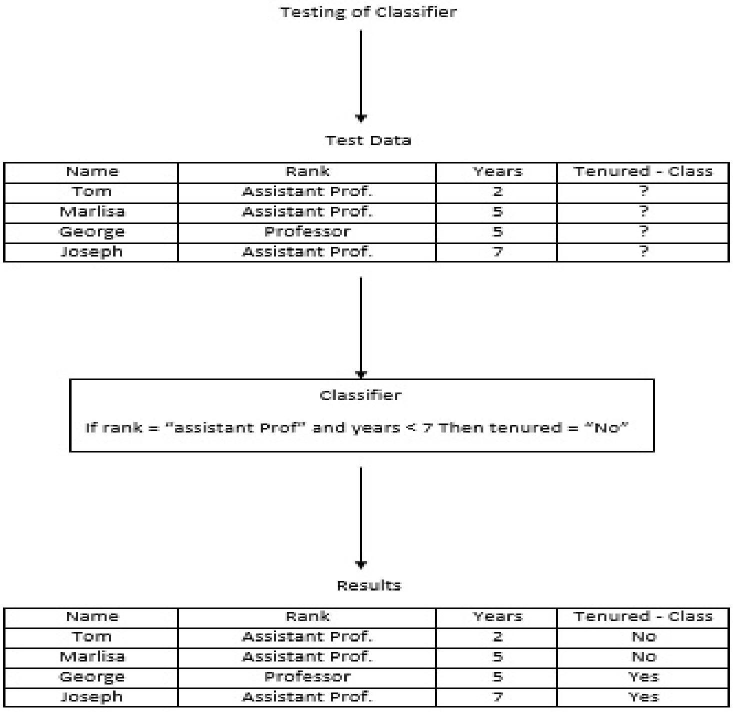
Test Data

Name	Rank	Years	Tenured - Class
Tom	Assistant Prof.	2	No
Marlisa	Assistant Prof.	5	No
George	Professor	5	Yes
Joseph	Assistant Prof.	7	Yes

Classifier

If rank = "assistant Prof" and years < 7 Then tenured = "No"

Fig. 11 Creating the pattern (Classifier).



Testing of Classifier

Test Data

Name	Rank	Years	Tenured - Class
Tom	Assistant Prof.	2	?
Marlisa	Assistant Prof.	5	?
George	Professor	5	?
Joseph	Assistant Prof.	7	?

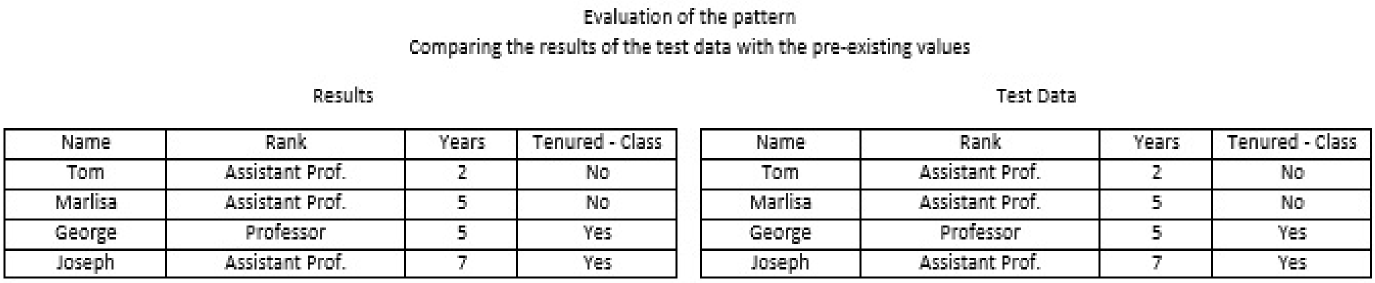
Classifier

If rank = "assistant Prof" and years < 7 Then tenured = "No"

Results

Name	Rank	Years	Tenured - Class
Tom	Assistant Prof.	2	No
Marlisa	Assistant Prof.	5	No
George	Professor	5	Yes
Joseph	Assistant Prof.	7	Yes

Fig. 12 Testing of Classifier.



Evaluation of the pattern

Comparing the results of the test data with the pre-existing values

Results

Name	Rank	Years	Tenured - Class
Tom	Assistant Prof.	2	No
Marlisa	Assistant Prof.	5	No
George	Professor	5	Yes
Joseph	Assistant Prof.	7	Yes

Test Data

Name	Rank	Years	Tenured - Class
Tom	Assistant Prof.	2	No
Marlisa	Assistant Prof.	5	No
George	Professor	5	Yes
Joseph	Assistant Prof.	7	Yes

Fig. 13 Evaluation of the Classifier.

4) Types of Classification Algorithms:-

a) Decision Tree Algorithm:

Decision Tree is an easy technique and more popular and commonly used. Decision Tree Algorithms are used for classification and prediction, requires little data preparation process, can handle both numeric and nominal data. Where decision tree algorithms help to reach certain decisions. There are two steps to build decision tree, these steps are Induction and Pruning. The decision tree is built into the induction step, while the numerous complexities of the tree are removed in the pruning step. The most common types of decision tree algorithm are C4.5, ID3, CART, and J48 [1,2].

To create the decision tree, the following are used:-

Entropy, it is a standard of the randomness in the processed information. Entropy equation is:-

$$E(D,X) = \sum P(c) \times E(c) \tag{1}$$

Information Gain, it minimizes the required information to classify the classes, where it minimizes the number of needed tests to classify the required class. The attribute with the highest information gain is selected. The Information Gain equation is:-

$$Info\ (D) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{2}$$

Gain, it is the minimization of required information according to X value. The attribute with the highest information gain is chosen as "best". Gain equation is:-

$$Gain(D,X) = Info(D) - E(D,X) \tag{3}$$

Gain Ratio, it splits the training dataset into parts, and considers the number of classes of the result with respect to the total classes. The attribute with the max gain ratio is used as a splitting attribute. Gain ratio equation is:-

$$Gain\ Rati(A) = \frac{Gain(A)}{Split\_Info(D)} \tag{4}$$

Where:-

$$Split\_Info(D) = -\sum_{j=1}^{m} \frac{D_j}{D} \times \log_2 \frac{D_j}{D}$$

(5)

Gini Index, it is calculated for binary variables only. It standardizes the impurity in training classes of dataset D. The Gini Index equation is:-

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{6}$$

b) K-Nearest Neighbor (KNN) Algorithm:

K-Nearest Neighbors (KNN) is a simple algorithm, and used for classification and regression, where it stores all available cases to classify the new cases by a most vote of its k-neighbors. The case assigned to the class is most common amongst its K-nearest neighbors measured by a distance function (Euclidean, Manhattan, Minkowski, and Hamming). The distance functions (Euclidean, Manhattan, Minkowski, and Hamming) are used for continuous variables, where the Hamming distance function is used for nominal data [1,2].

$$dist(X1,X2) = \sqrt{\sum_{i=1}^{n} (X1_i - X2_i)^2} \tag{7}$$

c) Neural Networks (NN) Algorithm:

Neural Networks (NN) is a mathematical algorithm that consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. The most common types of neural network algorithm are gradient descent, evolutionary algorithm, genetic algorithm [1, 2]. Neural Networks (NN) are used for classification, clustering and prediction, where neural network algorithm calculates the weights the neural network connected (including repeated iteration or cumulative calculation). The Neural network algorithm is divided into three types (Feed-forward networks, Feedback network, and Self-organization networks). Feed-forward networks are the perception back-propagation model and the function network as representatives, and mainly used in the areas such as prediction and pattern recognition. Feedback network is Hopfield discrete model and continuous model as representatives, and mainly used for associative memory and optimization calculation. Self-organization network is adaptive resonance theory (ART) model and Kohonen model as representatives, and mainly used for cluster analysis [9].

d) Naïve Bayes Algorithm:

Naïve Bayes Algorithm is a classification algorithm, is quick algorithm to predict the class of the dataset, and is based on Bayes' theorem which is an assumption of independence between predictors, and requires a small amount of training data to evaluation the necessary parameters. Naïve Bayes Algorithm is simple and especially effective in case of huge data sets. Naïve Bayes Algorithm is commonly used in text classification and with problems having multiple classes [1, 2]. Bayes' theorem is:-

$$P(H|X) = \frac{P(H|X)\times P(H)}{P(X)} \tag{8}$$

e) Support Vector Machine (SVM) Algorithm:

Support Vector Machine (SVM) Algorithm is most common algorithm, and is used for classification and regression analysis. The purpose of SVM is to find the best classification model to distinguish between elements of the two classes in the training data. The standard of the best classification model can be realized geometrically [1, 2].

f) Random Forest Algorithm:

Random Forest Algorithm is one of the most used algorithms. It uses for classification, regression and other tasks that is performed by the decision trees, where It is a meta-estimator that fits number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement [1, 2].

5.2 Regression Algorithms:

Regression algorithms are based on regression functions, where predict the output values based on input features from the data fed in the system. Regression algorithms estimate the relationships between the dependent variables and one or more independent variables or predictors. There are three main uses for regression algorithms (determining the strength of predictors, forecasting an effect, and trend forecasting). The most common uses for regression algorithms are trend analysis, time series prediction, financial forecasting, marketing, environmental modeling and drug response modeling. Classification and regression algorithms are used in prediction, but classification assigns data into discrete categories, while regression is used to predict a numeric or continuous value [1, 2, 10].

The most common types of regression algorithms are as following:-

a) Linear Regression Algorithm:

Linear Regression Algorithm is a linear method for modeling the relationship between the scalar response or criterion, and explanatory variables or the multiple predictors. Linear regression algorithm focuses on the conditional probability distribution of the response given the values of the predictors. Linear regression algorithm is a model that supposes a linear relationship between the input variables (x) and the single output variable (y). Linear regression algorithm is based on the following equation:-

$$Y = \beta_0 + \beta_1 X \tag{9}$$

The most popular applications of linear regression algorithm are in real estate predictions, salary forecasting, financial prediction, traffic.

b) Logistic Regression Algorithm:

Logistic Regression Algorithm is used in case the dependent variable is dichotomous. It is a form of binomial regression. Logistic regression is used to deal with data that has two possible criterions and the relationship between the criterions and the predictors. Logistic regression algorithm is based on the following equation:-

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{10}$$

The most popular applications of logistic regression algorithm are in industry that is applied through credit card scoring, fraud detection, and clinical trials.

c) Lasso Regression Algorithm:

Lasso regression algorithm performs variable selection and regularization, where it chooses only a subset of the provided covariates for use in the final model. The purpose of lasso regression algorithm is to get the subset of predictors that reduce prediction error for a quantitative response variable. The lasso regression algorithm works by imposing a constraint on the model parameters that leads regression coefficients for some variables to shrink toward a zero. Lasso regression algorithm is based on the following equation:-

$$\boldsymbol{N^{-1} \sum_{i=1}^{N} f(x_i\ , y_i\ , \alpha, \beta)} \tag{11}$$

The most popular applications of lasso regression algorithm are in financial networks and economics, and stock market forecasting.

5.3 Association Rules Algorithms:

1) What is Association Rules Algorithm:

Association Rules Algorithm is one of the most important data mining algorithms, where it depends on frequent mining of data, and aims to find interesting relations between variables in large dataset which meet minimum support (*min_sup*) and confidence level (*min*_conf) pre-determined by the user to discover a pattern which describes strongly associated relations in this dataset. Support refers to the number of occurrences of the items appear in the data, while Confidence refers to the number of times are found true. There are two-step for mining Association Rules (Frequent Itemset Generation, and Rule Generation) [11].

2) Applications of Association Rule Algorithms:

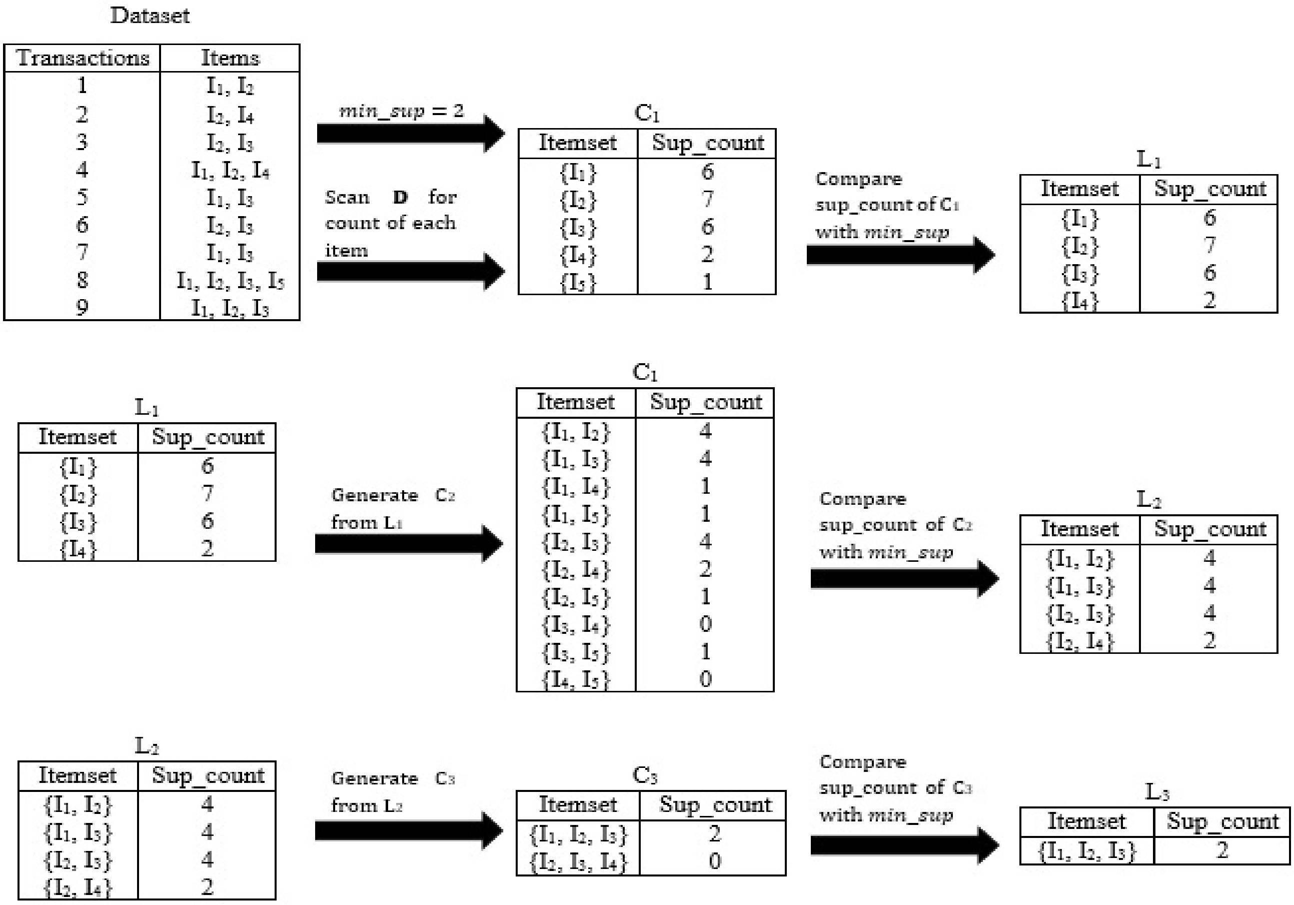
Association Rule Algorithms are used in many of fields (Market Basket Analysis, Medical Diagnosis, Census Data, and Protein Sequence). Market Basket Analysis is the most common example of association algorithms, where data is collected using barcode scanners in most supermarkets. Association rule algorithms are used in medical diagnosis through determining the likelihood of disease occurring in relation to various factors and symptoms. By Census Data, association rule algorithms can help governments to plan public services (such as health, education, transport and etc.) and help public businesses for build up new manufactories, shopping malls, and marketing particular products. Association rule algorithms helpful to discover association rules for sequence of twenty types of amino acids of protein during the synthesis of artificial proteins [12].

3) Mechanism of Association rules algorithms:

Association rules are created by searching data for frequent itemsets depending on the support and confidence factors. The form **A⇒B** is expression of association rule, where **A** and **B** are disjoint itemsets (**A∩B=∅**). For example, **I= {I$_1$, I$_2$, ..., I$_m$}** is set of binary attributes called items. **D** is a dataset which contains transactional records. Each transaction **T** is a non-empty itemsets (**T⊆I**). Each transaction **t** is referred as a binary vector. Suppose A and B are an itemsets. **A** and **B**, transaction **T** is said to contain **A** and **B** if and only if **A⊆T** and **B⊆T**. The form **A⇒B** is expression of association rule, where **A⊂I**, **B⊂I**, **A≠∅**, **B≠∅**, and **A∩B=∅**. The rule **A⇒B** holds in the transaction set D with support (*min_sup*), where ***min_sup*** is the percentage of transactions in D that contain **A∩B**. The rule **A⇒B** has confidence (***min_conf***) in the transaction set D, where *min*_conf is the percentage of transactions in **D** containing **A** that also contain B [11].

4) Example:-

Figure (14) illustrates an example for mechanism of association rules algorithm.


min_sup = 2
Scan D for count of each item
Compare sup_count of C1 with min_sup
Generate C2 from L1
Compare sup_count of C2 with min_sup
Generate C3 from L2
Compare sup_count of C3 with min_sup

Dataset

Transactions	Items
1	I_1, I_2
2	I_2, I_4
3	I_2, I_3
4	I_1, I_2, I_4
5	I_1, I_3
6	I_2, I_3
7	I_1, I_3
8	I_1, I_2, I_3, I_5
9	I_1, I_2, I_3

C_1

Itemset	Sup_count
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	1

L_1

Itemset	Sup_count
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2

L_1

Itemset	Sup_count
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2

C_1

Itemset	Sup_count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	1
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	1
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1
$\{I_4, I_5\}$	0

L_2

Itemset	Sup_count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2

L_2

Itemset	Sup_count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2

C_3

Itemset	Sup_count
$\{I_1, I_2, I_3\}$	2
$\{I_2, I_3, I_4\}$	0

L_3

Itemset	Sup_count
$\{I_1, I_2, I_3\}$	2

Fig. 14 Example of association rules algorithm.

5) Types of Association Rules Algorithms:

The most common types of association rules algorithms are as following:-

a) AIS Algorithm:

Artificial Immune Systems (AIS) algorithm is the first algorithm that was published and developed to generate all large itemset in a transaction dataset. It aims to discover qualitative rules, where it is only restricted to one item in the result. Its form is **X⇒Ij | α**. Where **X** refers to itemsets; **Ij** refers to a single item in the rang **I**; and **α** refers to confidence of the rule. AIS algorithm' results depend on the factor (***min_sup***) [13].

b) SETM Algorithm:

SETM Algorithm was suggested to be used in SQL to calculate large itemsets, where each item of the large itemsets **Lk** is in the form **<TID, itemset>**, **TID** is the ID of a k transaction; each item of the of candidate itemsets **Ck** is in the form **<TID, k itemset>** [13].

c) Apriori Algorithm:

Apriori algorithm is an easy and best association rules algorithm, and most common. Apriori is an algorithm for mining frequent items for boolean association rule. It uses a bottom-up method, and designed for finding association rules in the dataset that contains transactions, where it uses large itemset. Apriori algorithm' results depend on the factors **(*min_sup*,** and ***min*_conf)**. Any subset of a frequent itemset is considered as a frequent itemset by apriori algorithm, where the number of candidates is reduced by apriori algorithm. Apriori algorithm depend iterative method, where **k-itemsets** are used to find **(k+1)-itemsets** [13].

d) Frequent Pattern Growth (FP-Growth) Algorithm:

Manuscript received July 5, 2021
Manuscript revised July 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.7.26

FP growth algorithm is development of apriori algorithm, and faster than apriori algorithm. FP growth algorithm symbolizes frequent items in frequent pattern trees or so-called FP-tree. FP growth algorithm is used to find frequent itemset in a transaction dataset without candidate generation [14].

5.3 Clustering Algorithms:

1) What is Clustering Algorithm:

A cluster is a subset of objects which are similar. Clustering Algorithm is one of common data mining algorithms, and used for descriptive. It depends on dividing the dataset into sub-datasets (clusters), where the sub-datasets (clusters) are used to organize the objects in a way that makes each object inside the cluster similar to each other yet they differ from other clusters. Clustering algorithm requires the user specifies number of expected clusters. [15, 16] Cluster analysis are clustering **C = {C1, ..., Ct}** is a sub-dataset of the dataset **A**, the **P(A)** called partitions of the dataset **A** into **c** disjoint sub-datasets, covering the whole dataset.

$$A = \{x_1, x_2, \dots, x_k\} = \cup_{i=1}^{k} C_i$$

$$and\ C_i \cap C_j = \emptyset\ \ for\ all\ \ 1 \le i,\ \ j \le k \qquad (12)$$

Clustering algorithm divides a numeric data (A) by divide the values of A into clusters. Clustering takes the distribution of A into consideration, as well as the closeness of data points, and therefore is able to produce high-quality discretization results [2]. Clustering Algorithm applies for these applications such as economic science, web mining, and image processing.

2) Clustering Algorithms Methods:

a) Hierarchical methods:

In this method the dataset is divided into N clusters as a hierarchy or tree of clusters. Every cluster node contains child clusters; sibling clusters divide the points covered by their common parent [15, 16].

b) Density based methods:

In this method the clusters are considered as dense regions in the data space separated by sparse regions, where density function (such as mixture model) is used by this method. [15]

c) Partitioning Method:

In this method the dataset is divided into N clusters, and each cluster is represented by a centroid or a cluster representative, where that depends on the type of the object that is being clustered. If there are large number of clusters, the centroid can be further cluster to produces hierarchy within a dataset [15].

3) Types of Clustering Algorithms:

a) k-means algorithm:

k-means algorithm is an easy algorithm. It is used to divide large datasets. k-means algorithm is used with numeric data. It depends on partitioning methods. K refers to number of objects. k-means algorithm assigns the object to its nearest cluster center using Euclidean distance or Manhattan distance [2].

Euclidean distance:-

$$d(P,Q) = \sqrt{\left(\sum_{j=1}^{P}(x_j(P) - x_j(Q))^2\right)} \qquad (13)$$

Manhattan distance:-

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{iP} - x_{jP}|$$

(14)

b) k-medoids algorithm:

In k-medoids algorithm, each cluster is represented by one of the objects located near the center of the cluster. The iterative process of replacing representative objects by no representative objects continues as long as the value of the resulting clustering is improved. This value is estimated using the cost function that measures the average dissimilarity between an object and the representative object of its cluster [16].

c) The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH):

BIRCH is an algorithm depends on hierarchical clustering. BIRCH is executed on large datasets with high dimensions. BIRCH creates clustering tree which summarizes data by accumulating its zero, first, and second moments of the cluster [15].

6. CONCLUSIONS

Data mining is an exciting, important, Worthwhile, interesting and continued growth field. So, this paper presented overview of data mining from beginning to end. Data Mining has relationship with many of fields such as pattern recognition, artificial intelligence (AI), machine learning, databases, mathematical modeling, statistics, and management science & information systems. Data mining has several steps and processes. Therefore, this paper presented these steps beginning with data collection and end with evaluation of pattern. There are many data mining tools that can be used to implement algorithms, especially for beginners in this field. This paper provided the most common data mining tools, as well as the most popular programming languages that are used to implement and develop data mining algorithms. I prefer the WEKA tool for who want to implement data mining tools, because WEKA tool is an easy to use and handling, while who want to use programming languages, I prefer the Python language for the abundance of libraries in it. There are many of different and multi-tasking data mining algorithms. The most common data mining algorithms were mentioned as overview in this paper such as decision tree, k-nearest neighbors, neural networks, Naïve Bayes, support vector machine, random forest, regression, AIS, SETM, Apriori, FP-Growth, and k-means algorithms.

Acknowledgments

The author declares that no direct funding and financial interests in relation to this study. The author thanks the Chinese CSC scholarships foundation and Hefei University of Technology to give the opportunity to continue his higher education through a scholarship.

References

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining Book. Pearson Education, Inc. Aug 2011. ISBN: 978-7-111-31670-1.

[2] Jiawei Han, Micheline Kamber, Jian Pei. Data Mining Concepts and Techniques Book (Third Edition). Elsevier Inc. Mar 2012. ISBN: 978-7-111-37431-2.

[3] Wikipedia. Data mining. Wikipedia. Dec 2019. Available at: https://en.wikipedia.org/wiki/Data_mining.

[4] Roshan Ragel. Difference Between KDD and Data Mining. May 2011. Available at: https://www.differencebetween.com/difference-between-kdd-and-vs-data-mining/

[5] Robert Grünwald. Data Mining Tools. Novustat. Dec 2019. Available at: https://novustat.com/statistik-blog/data-mining-tools.html

[6] Wikipedia. Oracle Data Mining. Wikipedia. May 2020. Available at: https://en.wikipedia.org/wiki/Oracle_Data_Mining

[7] Kitty Gupta. The Best Programming Languages for Data Mining. FreelancingGig. Aug 2017. Available at: https://www.freelancinggig.com/blog/2017/08/07/best-programming-languages-data-mining/

[8] Wikipedia. MATLAB. Wikipedia. May 2020. Available at: https://en.wikipedia.org/wiki/MATLAB

[9] Priyanka Gaur. Neural Networks in Data Mining. International Journal of Electronics and Computer Science Engineering. 2012: Volume 1, Issue 3. 1449:1453.

[10] Nanhay Singh, Ram Shringar Raw, Chauhan R.K. Data Mining with Regression Technique. Journal of Information Systems and Communication. 2012: Volume 3, Issue 1. pp: 199-202.

[11] Mutlu Yüksel Avcilar, Emre Yakut. Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store. Canadian Social Science. 2014: Volume 10, Issue 3. pp: 75-83.

[12] Abhinav Rai. An Overview of Association Rule Mining & its Applications. UpGrad blog. Jun 2019. Available at: https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/

[13] Margaret H. Dunham, Yongqiao Xiao, Le Gruenwald, Zahid Hossain. A Survey of Association Rules. The Pennsylvania State University. Jan 2001. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.1602

[14] Yi Zeng, Shiqun Yin, Jiangyue Liu, and Miao Zhang. Research of Improved FP-Growth Algorithm in Association Rules Mining. Scientific Programming. 2015(3). pp:1-6.

[15] C.Anuradha, T.Velmurugan, R. Anandavally. Clustering Algorithms in Educational Data Mining: A Review. International Journal of Power Control and Computation (IJPCSC). 2015: Volume 7, Issue 1. pp: 47-52.

[16] M. Arumaiselvam, R. Anitajesi. Study of Clustering Methods in Data Mining. International Journal of Data Mining Techniques and Applications. 2018: Volume 7, Issue 1. pp: 55-59.

Abdulkawi Yahya Radman Al-Shamiri received his BSc degree in Computer Science from Hodeidah University, in 2007. From 2008 to 2016, he was Lecturer at Hodeidah University, The National University, and Community College for Medical Science & Technology, in Yemen. He has two books (Introduction for Computer Science, and Electronic management). Currently, he is a master degree student in “Artificial intelligence and pattern recognition” in Computer Application Technology at School of Computer and Information at Hefei University of Technology (HFUT). His research interests are in data mining and knowledge engineering.