Review Paper on Data Mining: Applications, Techniques and Algorithms

Fideline Kubwayo

Abstract— Data mining is a process of extracting some useful knowledge from a large amount of data. In this paper, we discussed a few of the data mining techniques, algorithms, and applications that are used by some of the organizations which have adapted data mining technology to improve their businesses and found excellent results.

Keywords— Data mining Applications, Data mining Techniques, Data Mining Algorithms.

1 INTRODUCTION

The development of Information Technology has generated a large number of databases and huge data in various areas and this has given rise to the concept of "Data mining" which is considered as the process of extraction hidden knowledge from large volumes of raw data. Data mining is also called a knowledge discovery process that extracts our data and presenting it in a form that is easily understood by humans [1].

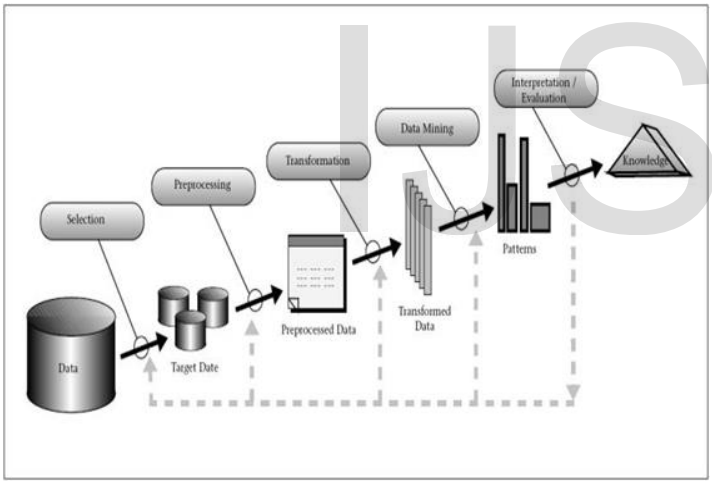

Figure1. Data mining process

Knowledge Discovery in Database or Data mining is a logical process that is used to search through a large amount of data to find useful data. The goal of this technique is to find previously unknown patterns.

Fideline Kubwayo is currently pursuing masters degree program in the department of Information Technology in University of Lay Adventists of Kigali, Rwanda, PH-(+250)783462771. E-mail: fidelinekubwayo1@gmail.com

Once these patterns are found they can further be used to make certain decisions for the development of their businesses. This process is divided into three main categories: Exploration, Pattern identification, Deployment, but few steps are involved in the process. Data cleaning- It is also known as the data cleansing, it is a phase in which noise data and relevant data removed from collection.

Data integration- At this stage, multiple data sources, often heterogeneous, may be combined in a common source. Data selection-At this step, the data relevant to the analysis is decided on and retrieved from the data collection. Data transformation- It is also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure. Data mining-It is the crucial step in which clever techniques are applied to extract patterns potentially useful. Pattern evaluation- In this step, strictly interesting patterns representing knowledge are identified based on given measures. Knowledge representation: It is the final phase in which the discovered knowledge is visually represented to the user. This essential set uses visualization techniques to help users understand and interpret data mining results [2].

2 DATA MINING APPLICATION

Data mining is widely used in diverse areas. There are several commercial data mining systems available today. Different organizations have been adopting data mining to their business processes to gain competitive advantages and help business grows. Here are some of the fields where data mining is applied − Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Intrusion Detection, Customer Relationship Management and Other Scientific Applications[3].

3 DATA MINING TECHNIQUES

There are several major data mining techniques have been developing and using in data mining projects recently. The techniques are classified into two main categories: Descriptive techniques are used to mine data and provide the latest information on past or recent events while the predictive techniques provide answers to the future queries that move

across using historical data as the chief principle for decisions. In this paper, we are going to discuss one descriptive technique (Clustering) and one predictive technique (Regression)[4].

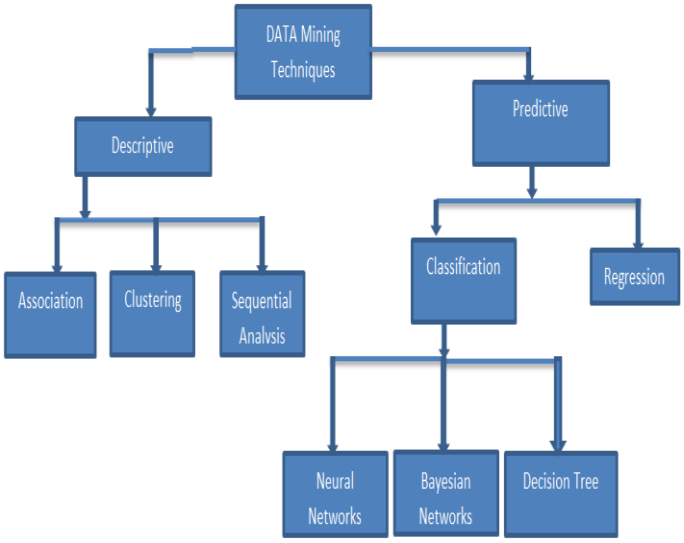

Figure2: Data Mining Techniques

1. Clustering Technique

In general when we say clustering in data mining we mean; a group of abstract objects into classes of similar objects is made and we treat a cluster of data objects as one group and objects in a group will be like[9] one another and different from the objects in other groups. This will help to understand the differences and similarities between the data. This is sometimes called segmentation and helps the users to understand what is going on within the database. For example, an insurance company can group its customers based on their income, age, nature of policy and type of claims.

There are six main methods of data clustering – the partitioning method, hierarchical method, density-based method, grid-based method, the model-based method, and the constraint-based method. Each method groups the data in a different way. In the density-based method, for instance, the data is clustered together according to its density, as the name suggests. In the grid-based method, the objects are organized to create a grid structure.

1.1 Benefits of Clustering Technique

When it comes to business, data mining is most commonly used by companies with a strong focus on customers – so retail, finance, and marketing are some of the key organizations that benefit from data mining. Data mining is so important to these kinds of businesses because it allows them to 'drill-down' into the data, and using clustering methods to analyze the data can help them gain further insights from the data they have on file. From this, they can examine the relationships between internal factors – pricing, product positioning, staff skills – and external factors – such as competition and the demographics of customers. For instance, utilizing one of the clustering methods during data mining can help the business to identify distinct groups within their customer base. They can cluster different customer types into one group based on different factors, such as purchasing patterns. The factors analyzed through clustering can have a big impact on sales and customer satisfaction, making it an invaluable tool to boost revenue, cut costs, or sometimes even both.

1.2 Disadvantages of Clustering Technique

Among the disadvantages, we can mention The big data sets, which make useless the key concept of clustering, the distance between observations thanks to the curse of dimensionality. Also, we see the deficiencies of existing algorithms to mean:

- Inability to rank the variables by contribution in the heterogeneity of dataset.
- Inability to detect special patterns - cores, kernels, borders of clusters, the zone of mixing, noise, and outliers.
- Inability to estimate the correct number of clusters.
- Each algorithm is intended to find the same type of clusters and fails if we have a mixture of clusters with different characteristics.

2. Regression Technique

Regression is one of the data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables. Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends.

2.1 Types of Regression Techniques

The simplest and oldest form of regression is linear regression used to estimate a relationship between two variables. This technique uses the mathematical formula of a straight line (y = MX + b). In plain terms, this simply means that, given a graph with a Y and an X-axis, the relationship between X and Y is a straight line with few outliers. For example, we might assume that, given an increase in population, food production would increase at the same rate — this requires a strong, linear relationship between the two figures. To visualize this, consider a graph in which the Y-axis tracks population increase, and the X-axis tracks food production. As the Y value increases, the X value would increase at the same rate, making the relationship between them a straight line. Advanced techniques, such as multiple regressions, predict a relationship between multiple variables — for example, is there a correlation between income, education and where one chooses to live? The addition of more variables considerably increases the complexity of the prediction. There are several types of multiple regression techniques including standard, hierarchical, setwise and stepwise, each with its application.

- **Standard multiple regression** considers all predictor variables at the same time. For example 1) what is the relationship between income and education (predictors) and choice of the neighborhood (predicted), and 2) to what degree does each of the individual predictors contribute to that relationship?
- **Stepwise multiple regression** answers an entirely different question. A stepwise regression algorithm will analyze which predictors are best used to predict the choice of the neighborhood — meaning that the stepwise model evaluates the order of importance of the predictor variables and then selects a relevant subset. This type of regression problem uses "steps" to develop the regression equation. Given this type of regression, all predictors may not even appear in the final regression equation.
- **Hierarchical regression**, like stepwise, is a sequential process, but the predictor variables are entered into the model in a pre-specified order defined in advance, i.e. the algorithm does not contain a built-in set of equations for determining the order in which to enter the predictors. This is used most often when the individual creating the regression equation has expert knowledge of the field.
- **Setwise regression** is also similar to stepwise but analyzes sets of variables rather than individual variables.

3. Clustering and Regression Algorithms

3.1 Clustering Algorithms

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering has many algorithms but let's talk about five common clustering algorithms.

K-Means Clustering

K-Means is probably the most well-known clustering algorithm. It's taught in a lot of introductory data science and machine learning classes. It's easy to understand and implement in code! Check out the graphic below for an illustration. K-Means has the advantage that it's pretty fast, as all we're doing is computing the distances between points and group centers; very few computations! It thus has a linear complexity O (n). On the other hand, K-Means has a couple of disadvantages. Firstly, you have to select how many groups/classes there are. This isn't always trivial and ideally, with a clustering algorithm, we'd want it to figure those out for us because the point of it is to gain some insight from the data. K-means also start with a random choice of cluster centers and therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency. Other cluster methods are more consistent.

Mean-Shift Clustering

Mean shift clustering is a sliding-window-based algorithm that attempts to find dense areas of data points. It is a centroid-based algorithm meaning that the goal is to locate the center points of each group/class, which works by updating candidates for center points to be the mean of the points within the sliding-window. These candidate windows are then filtered in a post-processing stage to eliminate near-duplicates, forming the final set of center points and their corresponding groups. In contrast to K-means clustering, there is no need to select the number of clusters as mean-shift automatically discovers this. That's a massive advantage. The fact that the cluster centers converge towards the points of maximum density is also quite desirable as it is quite intuitive to understand and fits well in a naturally data-driven sense. The drawback is that the selection of the window size/radius "r" can be non-trivial.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a density-based clustered algorithm similar to mean-shift, but with a couple of notable advantages. DBSCAN poses some great advantages over other clustering algorithms. Firstly, it does not require a pe-set number of clusters at all. It also identifies outliers as noises, unlike mean-shift which simply throws them into a cluster even if the data point is very different. Additionally, it can find arbitrarily sized and arbitrarily shaped clusters quite well. The main drawback of DBSCAN is that it doesn't perform as well as others when the clusters are of varying density. This is because the setting of the distance threshold ε and pinpoints for identifying the neighborhood points will vary from cluster to cluster when the density varies. This drawback also occurs with very high-dimensional data since again the distance threshold ε becomes challenging to estimate.

Expectation-Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

There are 2 key advantages to using GMMs. Firstly GMMs are a lot more flexible in terms of cluster covariance than K-Means; due to the standard deviation parameter, the clusters can take on any ellipse shape, rather than being restricted to circles. K-Means is a special case of GMM in which each cluster's covariance along all dimensions approaches 0. Secondly, since GMMs use probabilities, they can have multiple clusters per data point. So if a data point is in the middle of two overlapping clusters, we can simply define its class by saying it belongs X-percent to class 1 and Y-percent to class 2. I.e. GMMs support mixed membership.

Agglomerative Hierarchical Clustering

Hierarchical clustering algorithms fall into 2 categories: top-down

or bottom-up. Bottom-up algorithms treat each data point as a single cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all data points. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC. This hierarchy of clusters is represented as a tree. The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. Hierarchical clustering does not require us to specify the number of clusters and we can even select which number of clusters looks best since we are building a tree. Additionally, the algorithm is not sensitive to the choice of distance metric; all of them tend to work equally well whereas, with other clustering algorithms, the choice of distance metric is critical. A particularly good use case of hierarchical clustering methods is when the underlying data has a hierarchical structure and you want to recover the hierarchy; other clustering algorithms can't do this. These advantages of hierarchical clustering come at the cost of lower efficiency, as it has a time complexity of O (n^3), unlike the linear complexity of K-Means and GMM.

3.2 Regression Algorithms

Regression algorithms fall under the family of Supervised Machine Learning algorithms which is a subset of machine learning algorithms [5]. One of the main features of supervised learning algorithms is that they model dependencies and relationships between the target output and input features to predict the value for new data. Regression algorithms predict the output values based on input features from the data fed in the system.

1. Simple Linear Regression model: Simple linear regression is a statistical method that enables users to summarize and study relationships between two continuous (quantitative) variables. Some of the most popular applications of linear regression algorithms are in financial portfolio prediction, salary forecasting, and real estate predictions and traffic in arriving at ETAs.

2. Lasso Regression: LASSO stands for Least Absolute Selection Shrinkage Operator wherein shrinkage is defined as a constraint on parameters. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The algorithm operates by imposing a constraint on the model parameters that cause regression coefficients for some variables to shrink toward a zero. Lasso regression algorithms have been widely used in financial networks and economics.

3. Logistic regression: One of the most commonly used regression techniques in the industry which are extensively applied across fraud detection, credit card scoring and clinical trials, wherever the response is binary has a major advantage. One of the major upsides is of this popular algorithm is that one can include more than one dependent variable which can be continuous or dichotomous. The other major advantage of this supervised machine learning algorithm is that it provides a quantified value to measure the strength of association according to the rest of the variables. Despite its popularity, researchers have drawn out its limitations, citing a lack of robust technique and also a great model dependency. Today enterprises deploy Logistic Regression to predict house values in real estate business, customer lifetime value in the insurance sector and are leveraged to produce a continuous outcome such as whether a customer can buy/will buy scenario.

4. Multiple Regression Algorithms: This regression algorithm has several applications across the industry for product pricing, real estate pricing, marketing departments to find out the impact of campaigns. Unlike the linear regression technique, multiple regressions are a broader class of regressions that encompasses linear and nonlinear regressions with multiple explanatory variables. Some of the business applications of multiple regression algorithms in the industry are in social science research, behavioral analysis and even in the insurance industry to determine claim worthiness.

5. Multivariate Regression algorithm: This technique is used when there is more than one predictor variable in a multivariate regression model and the model is called a multivariate multiple regression. Termed as one of the simplest supervised machine learning algorithms by researchers, this regression algorithm is used to predict the response variable for a set of explanatory variables. Industry application of the Multivariate Regression algorithm is seen heavily in the retail sector where customers choose several variables such as brand, price, and product. The multivariate analysis helps decision-makers to find the best combination of factors to increase footfalls in the store.

4 Conclusion

The purpose of this paper was to present a detailed description of data mining applications, techniques, and algorithms. Data mining is very important regarding finding patterns, discovering interesting knowledge from large amounts of data in different business domains. The various techniques and algorithms used for mining data are discussed in detail. Data mining also has a wide range of applications in the most industry where data is the key factor. For this data mining is a promising concept in the development of information technology and also to get the best out of the data extraction.

References

[1] Prajapati. D, Prajapat. J, "Handling missing values: Application to University Data Set", August 2011.

[2] Grabmeier. J, Rudolph. A, "Technique of Clustering Algorithms in Data Mining", Data Mining and Knowledge Discovery, 2002.

[3] https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

[4] Jiawei. H, Micheline. K, "Data Mining Concepts and Techniques", 2006.

[5] https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/

[6] https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/