International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

A Comparative Study on Data Mining Algorithms for Classification & Regression

D. Kavitha*, K. Sivasankari, M. Pavethra
CSE, Anna University/Dhaanish Ahmed College of Engineering, Chennai, TamilNadu, India

ABSTRACT

Today, most of the organizations are actively collecting and storing data in large databases. The increasing demandforretrieval and analysis is answered by an efficient method called as "Data Mining" (DM). It is the process of extracting hidden information from large database/data warehouse. For the retrieval and analysis, DM uses different types of algorithms.Based on its applications,data mining algorithms are classified into five types such as, classification, regression, segmentation, association, sequence analysis.In this paper we present a comparative study among classification and regression algorithms. This paper providesa complete knowledge about the explained algorithms and a comparison between the algorithms presented in this section improves the value of this study.

Keywords : Data Mining, Classification, Regression, SVM, KNN.

I. INTRODUCTION

In an Information Technology (IT) world Data Mining is the process of extracting valuable, implicit, unknown and potentially useful information (such as knowledge rules, constraints, regularities)from data in databases and analyses the large amount of data to find new and hidden information [1].It is also referred as Knowledge Discovery in Databases (KDD), knowledge mining from databases, knowledge extraction, data archaeology, data dredging, data analysis, etc [2].Data mining is a significantapproach and user friendly in reading reports, reducing errors and controls the quality [3].Data mining can be used in various areas including sales/ marketing, banking, insurance, health care, transportation, medicine, business transactions, scientific purpose, surveillance video and pictures, satellite sensing, text reports and memos (e-mail messages) and World Wide Web (WWW) repositories, etc[4].The main objective of the data mining process is to extract information from a set of data and convert it into an understandable structure for future use [5].

Data mining (KDD) process consist of five steps namely, data selection, pre-processing (or) data cleaning, data transformation (or) reduction, data mining task (or) algorithm selection, and finally post processing (or) explanation of discovered knowledge. The important task of data mining is data pre-processing [6]. Essentially, data mining can be classified into several types based on models (or) patterns and the discovered knowledge. Based on whether they seek to build models or to find patterns DM can be classified as building models, and the second type is pattern detection [7]. Based on the discovered knowledge DM can be broadly divided into supervised learning and unsupervised learning, where the former requires the data to be pre-classified and in contrast, the latter does not require pre classification of the data [8].

Data mining utilizes two main forms such as, verification-driven and discovery-driven data mining, where in the former, the user suggests a hypothesis, and the system tries to validate it and in the latter,

the system automatically extracts new information from data, and forms the focus [9].DM understands the actions of identifying,validating and using for prediction, structural patterns in data, which is grouped into five categories: decision trees, classification rules, association rules,clusters and numeric prediction [10].

II. METHODS AND MATERIAL

Classification and regression are the two main purposes of DM algorithms. Classification is the process of classifying the known structure to apply new data. Regression is a statistical technique for investigating and modelling relationships between variables. Classification can be applied to both simple data (nominal, numerical, categorical and Boolean, etc) and to complex data (time series, graphs, trees etc).Regression analysis can be used in different applications such as engineering, chemical sciences, economics, biological sciences and other scientific fields. The main goal of classification is to predict the target class (Yes/ No). Regression technique is used to predict the range of numerical values [43]. The main purpose of regression analysis is to find a mathematical model to explain the relations between a response variable and the regressor variable.Some of the data mining algorithms used for both classification and regression are explained as follows:

2.1 K-Nearest Neighbour

K-Nearest Neighbor (KNN) is the type of supervised learning method. It is used for both pattern recognition and classification [11]. In KNN, a particular (test) tuple set is compared with the already available training data set which is similar to it [12]. It calculates the distance between the test data and the training data by using the Euclidean distance function. The output of the KNN classification is the class membership. Thus the KNN classification has two stages; the first is the determination of the nearest neighbors and the second is the determination of the class using those neighbors [13].The working process of KNN classifier is defined below[14]:

1. Calculate the distance between the attributes of training and test data sets.
2. Sort all the training datas based on the distance values.
3. Determine the neighbors (k) which are near to the test data.
4. Assign the majority class of training data to the test data.

The Euclidean distance between the training data set and the test data set can be derived as,

$$D(p,q)=\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+...+(p_n-q_n)^2} \quad (1)$$

$$D(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad (2)$$

Where, $P=\{p_1,p_2,...,p_n\}$ is the set of training data set, $Q=\{q_1,q_2,...,q_n\}$ is the set of test data set and D is the Euclidean distance.

In KNN classification, the class membership of the test data sets is as same as the class of training data sets which are nearer to the test data [15]. Let C be the class membership of KNN classification.The membership of test data set can be calculated by using the following expression.

$$C_i=\{p\in C_n;D(p,p_i)\le D(p,p_r),i\#r\} \quad (3)$$

Basically, the value of K (neighbor data) is an odd number such as, K = 1, 3, 5… So that we can avoid tie between the data sets.

Algorithm [16]:

Step 1: Set training data = $TT(p)$,

Step 2: Set testing data = $TS(q)$.

Step 3: For all $TT(p)$ and for all $TS(q)$,

Step 4: Compute the distance D by using the Eqn. (2).

Step 5: Find the smallest distance k.

Step 6: Locate the corresponding datas $TS(q_1),\ldots,TS(q_k)$.

Step 7: Assign $TT(p) = C_i$ //class with majority.

2. 2 Naive Bayes

Bayesian network is a directed, acyclic graph, which represents the set of random variables and their dependencies. The Bayesian classifier is constructed based on the bayesian network [17]. Naive Bayes classifier is a linear classifier, in which the attributes are considered as independent and have equal weight [18]. In naïve bayesclassifier all the attributes are independent of each other attribute which are given in the specific context of the class [19]. Naïve bayes classifier is a supervised learning method with probabilistic approach based on assumptions of the attributes independence. NB can be used in different areas such as, image classification, text classification, web mining, fraud detection etc [20]

The Bayes theorem describes the probability of an event. It is a powerful tool for making statistical interference [21]. In Bayes theorem, let E be the evidence and H be the hypothesis and $E=\{e_1,e_2,...,e_n\}$ be the set of samples with n attributes. $P(H/E)$ is the probability that the hypothesis H holds the given evidence E. $P(H/E)$ is the a posteriori probability of H conditioned on E and $P(E/H)$ is the a posteriori probability of E conditioned on H. $P(H)$ is the a priori probability of H, and $P(E)$ is the a priori probability of E [22]. It can be expressed as,

$$P(H/E)=\frac{P(E/H)\ P(H)}{P(E)} \quad (4)$$

Consider S be the training set of samples with their class, which consist of n - dimensional attribute vector and a set of predefined class $C=\{c_1,c_2,...,c_m\}$, which consist of m classes, and consider a set of samples $Y=\{y_1,y_2,...,y_n\}$. The classifier will assume that the sample Y belongs to the class which has the highest posteriori probability.To predict the class type of Y, calculate $P(Y/C_u)\ P(C_u)$ for each data. The classifier will predict the class of Y is C_u if and only if it is the majority class of training data set [23].

Algorithm

Step 1: Set $S=$ set of training datas,

$Y=\{y_1, y_2,\ldots,y_n\}$ = set of sample datas,

$C=\{c_1,c_2,\ldots,c_m\}$= set of class.

Step 2: If $P(C_u/Y)>P(C_v/Y)$ for $1\le v\le m,\ u\ne v$.

Step 3: Then $Y\in C$.

Step 4: Maximize posteriori hypothesis.

Step 5: By using Bayes theorem,

$$P(C_u/Y)=\frac{P(Y/C_u)\ P(C_u)}{P(Y)}$$

Step 6: Maximize $P(Y/C_u)\ P(C_u)$ and $P(Y)$ as constant.

Step 7: Calculate $P(Y/C_u)=\prod_{m=1}^{n} P(y_m/C_u)$ and

$$P(C_u)=\frac{freq(C_u/S)}{|S|},$$

Where

$$P(Y/C_u)=P(y_1/C_u)\times P(y_2/C_u)\times....\times P(y_n/C_u)$$

2. 3 Decision Tree Algorithm

Decision tree algorithm is an important approach for data mining methods. It is used for both classification and prediction. The decision tree is the flow chart like structure that isolates the set of relevant datas into an already defined class[24]. This algorithm takes the training datas as an input and produces the output as the decision tree for those training data sets.The basic approach of decision tree is greedy algorithm, which process the data in top-down manner of decision tree structure [25]. Decision trees are the most familiar type of the rule-based classifiers [26]. The construction of decision tree consists of the recursive partitioning of the training data set [27].

Consider a training set $S=\{(a_1,b_1),....,(a_n,b_n)\}$, where $\{a_1,...,a_n\}$ are the set of feature vectors and $\{b_1,...,b_n\}$ are the set of labels. This process is recursive. The nodes will stop growing until they reach the stopping criteria (SC). *BestSplit* returns

the best split point and *FindSplit* splits the data according to the *BestSplit* point [28].

Algorithm

Step 1: Initialize training sample S, attribute list L.
Step 2: Create node N.
Step 3: If $S(N) \in C$, return N as leaf node.
Step 4: If $(L = null)$,
return N as leaf node, and mark the node as belongs to large number category.
Step5: Select $A \in I$ (largest information gain) and set $N = A$.
Step 6: For each $a_i \in A$, partition N.
Step 7: If $a_i = A$, generate branch from N and set s_i = obtained sample set.
Step 8: If ($s_i = null$), mark leaf node with the most number of sample types
Step 9: Else mark it with return value $T(s_i, L - A)$
Step 10: Output the decision tree T.

The most widely used algorithm in decision trees is ID3 algorithm.

2.3.1 ID3 Algorithm:

ID3 stands for induction decision tree-version 3. ID3 uses a recursive manner to construct a decision tree from the given set of data [29]. It will be performed in top-down manner. ID3 algorithm only contains discrete data. ID3 algorithm computes the information gain of test attributes of non-leaf nodes in decision trees. It begins at the root and decides which character provides the better classification.It is a greedy technique that selects the sequential attributes based on the information gain associated with the attributes. The attribute which has the maximum information gain or greatest entropy reduction is selected as the test attribute for the next node [30].

Consider a data set X, and X_i be the number of datas [31]. Entropy is a function of information. It is calculated based on proportion of target values. It is a measure of uncertainty with a random variable. The values of entropy ranged from 0 to 1.Entropy is defined as, $E(X) = \sum_{i=1}^{n} - p_i \log_2 p_i$ (5)

Where n is the total number of different values and p_i is the proportion of tuples of the data set.Information gain is the method of selecting the attribute. The information gain of an attribute is defined as, $Gain(X, A) = E(X) - \sum_u \frac{|X_u|}{|X|} E(X)$ (6)

Algorithm

Step 1: Set A = set of attributes,
C = class attribute and
T = set of transactions.
Step 2: If $A = 0$, return leaf-node, with the class value with Transaction T.
Step 3: If T consist of transactions in class attribute, return leaf-node with C.
Step 4: Else,
(a) Consider the *best* classifier S.
(b) Let $S = \{s_1, \ldots, s_n\}$ and partition of $T = \{T(s_1), \ldots, T(s_n)\}$, where $T(s_i)$ has the attribute value of s_i.
(c) Return tree, whose root is S and edges are $s_1, \ldots, s_n$, for every i, theedge goes to ID3.

2.4. Neural Network (NN)

Neural Network (NN) is a mathematical (or) computational model based on biological neural networks. It is also defined as an imitation of biological neural system. It is also known as Artificial Neural Network (ANN)(or) Simulated Neural Network (SNN) [32]. It consists of an interconnected network of artificial neurons and process the information using a connectionist way to calculation. It is one of the supervised (or) associative learning in which the network is provided with a set of inputs and matching outputs. Deciding the ANN structure is a major design issue and can be critical for the classification performance [33]. The components of

ANN are: input layer, output layer, hidden layer, bias, net activation, activation function, network learning and stopping criterion. Neural Network based data mining consist of three main phases:

- ✓ Network construction & training: constructs and trains a three-layer neural network.
- ✓ Network Pruning: Aims to removing redundant links & units without increasing the classification error rate.
- ✓ Rule Extraction: Extracts the classification rules from the pruned network [34].

The neural networks are used to display the complex relationships between inputs and outputs or to find designs in the data.

Algorithm

Step 1: Set weight = w

Training set = T

Step 2: Repeat

For $(e \in T)$

assign $O = output(network, e)$

Forward pass

$T =$ output for e.

Calculate $Error = (T - O)$

Calculate Δwi for all weights from hidden to output layer

Calculate Δwi for all weights from input to hidden layer

Update w.

Step 3: Until stopping criterion satisfied.

Step 4: Return *network*

2.5. Support Vector Machines (SVM)

Support Vector Machine is a computer algorithm which learns by example to assign labels to objects [35]. It is a powerful tool for solving both classification and regression problems. It is one of the supervised learning method and one of the best-known classification methods [36]. SVMs are based on statistical learning theory, which is used to solve two-class (binary) problems without the loss of generality. The main goal of SVM classifiers is to determine the location of decision boundaries (hyper planes) which produces the optimal separation of classes [37]. Basically it is used to solve linear problems but now it is also extended to handle nonlinear decision problems. Some of the characteristics of SVM are good generalization performance, absence of local minima and sparse representation of solution. It is based on the Structural Risk Minimization (SRM) principle which minimizes the upper bound of the generalization error [38].

Consider a problem of classifying m points into n-dimensional real space R^N which can be represented as $m \times n$ matrix [39]. Consider a set of input samples $(a_x, b_x), x = 1, 2, ..., N$, where N is the number of samples, $a_x \in R^N$ and $b_x = \{+1, -1\}$ has two classes such as, positive class and negative class, i.e. $b_x = 1$ is the positive class and $b_x = -1$ is the negative class [40]. The classification hyper plane in N - dimensional space is $\omega\ a + z = 0$. Consider a hyper plane $f(X) = 0$, which separates the data

$$f(X) = \omega^T a + z = \sum_{y=1}^{N} \omega_y\ a_y + z = 0 \quad (7)$$

Where ω is a vector on N dimensional space and z is a scalar. $b_x\ f(X_x) = b_x\ (\omega^T\ a_x + z) \geq 1$, (8)

Where $x = 1, 2, ..., N$ The QP problem is expressed as [41],

$$\min\ \phi(\omega) = \frac{1}{2} \| \omega \|^2 \quad (9)$$

Where $b_x\ [\omega . a_x + z] \geq 1,\ x = 1, 2, ,..., N$

If the training data is not separated linearly, the formula must be modified to allow the classification violation samples as below:

$$\min\ \phi(\omega, \xi) = \frac{1}{2} \| \omega \|^2 + C . \left(\sum_{x=1}^{N} \xi_x \right) \quad (10)$$

Where $b_x\ [\omega . a_x + z] \geq 1 - \xi$, $x = 1, 2, 3, ... N$

$\xi_x \geq 0$, $x = 1, 2, 3, ... N$.

Introduce Lagrange multipliers, the dual formula for this problem can be written as,

$$\max W(\alpha) = \sum_{x=1}^{N} \alpha_x - \frac{1}{2} \sum_{x,y=1}^{N} \alpha_x \alpha_y b_x b_y\ (a_x, a_y) \quad (11)$$

Where $0 \le \alpha_x \le C$, $x = 1, 2, ..., N$

$$\sum_{x=1}^{N} b_x \alpha_x = 0 \quad (12)$$

By solving the above problem, the classifier can be expressed as, $f(X) = sign\left(\sum_{x=1}^{N} \alpha_x b_x (a \,.\, a_x) + y\right)$ (13)

Where α_x is the solution of QP problem [42].

Algorithm

Step 1: Set input $= (a_x, b_x)$ where $x = 1, 2, ..., N$, $a_x = R^n$ and $b_x = \{+1, -1\}$

Step 2: $f(X) = \omega^T a + z = \sum_{y=1}^{N} \omega_y a_y + z = 0$

Step 3: Minimize the QP problem as, $\min \phi(\omega, \xi) = \frac{1}{2} \| \omega \|^2 + C.\left(\sum_{x=1}^{N} \xi_x\right)$.

Step 4: Calculate the dualLagrangian multipliers

$$\min L(\omega, y, \alpha) = \frac{1}{2} \| \omega \|^2 - \sum_{x=1}^{N} \alpha_x b_x (\omega a_i + z) + \sum_{x=1}^{N} \alpha_x$$

Step 5: Calculate the dual quadratic optimization (QP) problem

$$\max W(\alpha) = \sum_{x=1}^{N} \alpha_x - \frac{1}{2} \sum_{x,y=1}^{N} \alpha_x \alpha_y b_x b_y (a_x, a_y).$$

Step 6: Solve the optimization problem $\sum_{x=1}^{N} b_x \alpha_x = 0$

Step 7: Output as $f(X) = sign\left(\sum_{x=1}^{N} \alpha_x b_x (a \,.\, a_x) + y\right)$

2.6 Simple Linear Regression Model

The simple linear regression analysis displays the relation between the independent (explanatory) and dependent (outcome) variable. The purpose of simple regression analysis is to evaluate the corresponding impact of a predictor variable on a particular outcome [44]. It contains only one independent variable X_i, where $i = 1, 2, ..., N$ and the corresponding dependent variable is labeled. It can be expressed as,

$$Y_i = a + b X_i + e_i \quad (14)$$

Where a is intercept, b is the slope of the regression line, and e_i is the random error term. The regression error is the vertical distance between the observed response and the true regression line.The main goal of the linear regression is to place a straight line through the data thatpredicts Y in view of X[45].

Algorithm

Step 1: Set Class models $C_i \in R^{y \times x_i}$, where $i = 1, 2, ..., N$ and Vector $v \in R^{y \times 1}$

Step 2: Evaluate $\hat{\beta}_i \in R^{x_1 \times 1}$ against each model where, $\hat{\beta}_i = \frac{C_i^T v}{(C_i^T C_i)}$, $i = 1, 2, ..., N$

Step 3: $\hat{v}_i$ is compared for each $\hat{\beta}_i$, such as $\hat{v}_i = C_i \hat{\beta}_i$, $i = 1, 2, ..., N$.

Step 4: Calculate the distance between the original and predicted response variables using the following expression,

$d_i(v) = \| v - \hat{v}_i \|_2$, $i = 1, 2, ..., N$

Step 5: Select the class with the minimum distance $d_i(v)$.

Step 6: Output class v.

2.7 LASSO Regression

LASSO can be defined as Least Absolute Shrinkage and Selection Operator, which combines the L_1 penalty into an OLS (Ordinary Least Squares) loss function. It shrinks some of the coefficients to zero [46].It is used for both estimation and variable selection. It minimizes the residual sum of square with respect to the sum of the coefficients which is less than a constant. It is used in subset selection and ridge regression and remembers the good features of both the methods [47].Consider a set of data (x^n, y_n), $n = 1, 2, 3, ... N$, where $x^n = (x_{n_1}, ..., x_{n_p})^T$ are predictor variables and y^n is the response variable. y^n S are conditionally independent on the given x_{nm} S. Standardize x_{nm}, so that we can get

$$\sum_{n} x_{nm} / N = 0 \quad (15)$$

$$\sum x^2_{nm} / N = 1. \quad (16)$$

Let $\hat{\beta}=(\hat{\beta}_1,\ldots,\hat{\beta}_p)^T$ and LASSO estimate as $(\hat{\alpha},\hat{\beta})$

$$(\hat{\alpha},\hat{\beta})=\ argmin\left\{\sum_{n=1}^{N}\left(y_n-\alpha-\sum_m \beta_m x_{nm}\right)^2\right\} \text{ subject to } \sum_m|\beta|\le t \cdot \quad (17)$$

where $t \geq 0$ is a tuning parameter. For all t the solution for α is $\hat{\alpha}=\bar{y}$. If $\bar{y}=0$, then omit α.

Consider $E=\{n;\ \delta_n^T\beta=t\}$ is an equality set, $S=\{n;\ \delta_n^T\beta<t\}$ is a slack set and $g(\beta)=\sum_{n=1}^{N}\left(y_n-\sum_m \beta_m x_{nm}\right)^2$, δ_n where $n=1,2,\ldots,2^p$. The condition $\sum|\beta_m|\le t$ and $\delta_n^T\beta\le t$ are equivalent to each other for all n. Consider M_E is a matrix which has δ_n rows for $n\in E$.

Algorithm

Step 1:Set $E=\{n_0\}$. //initialize E

Step 2: Set $\delta_{n_0}=sign(\hat{\beta}^0)$, where $\hat{\beta}_0=$ overall least squares.

Step 3: Find $\hat{\beta}$ to minimize $g(\beta)$ subject to $M_E\beta\le t$.

Step 4: While $\{\sum|\hat{\beta}_m|>t\}$

Step 5: Add n to the equality set E, where $\delta_n=sign(\hat{\beta})$.

Step 6: Find $\hat{\beta}$ to minimize $g(\beta)$ subject to $M_E\beta\le t$.

2.8 Logistic Regression

Logistic Regression (LR) is a linear algorithm, which is the process of relating dependent and independent variables using a logistic distribution functional form. The regression model can be formulated mathematically by relating the probability of some event [48]. It provides the linear relationship between the input and the output [49]. The logistic regression model calculates the class membership probability for one of different categories in the data set [50]. It is used to model the binary response data. If the response is binary, it typically takes the form of $0/1$ where 1 indicates the success and 0 indicates the failure [51].Consider data d, weights (a,b) and class label c .Assume the probability as,

$$P(c=\pm1|d,a)=\frac{1}{1+\exp(-c(a^T d+b))} \quad (18)$$

Calculate (a,b) by minimizing the negative log-likelihood. It can be expressed as [52],

$$\min_{a,b}\sum_{i=1}^{l}\log\left(1+e^{\left(-c_i/a^T d_i+b\right)}\right) \quad (19)$$

The algorithm for logistic regression can be expressed in [53] as,

Algorithm

Step 1: Set data = d and weights (a, b).

Step 2: Set class label = c.

Step 3: Calculate the probability by using Eqn.(18)

Step 4: Initialize training instance d_i and labels $c_i\in\{1,-1\}$, where $i=1,2,\ldots,l$.

Step 5: Calculate (a,b), by minimizing the negative log-likelihood as Eqn. (19),

Step 6: To obtain a simple derivation assume $d_i^T\leftarrow[d_i^T,1]$

Step 6: Calculate regularization abilities by adding regularization term $\frac{1}{2}a^T a$

Step 7: Output the regularized logistic regression,

$$\min_a f(a)\equiv\frac{1}{2}a^T a+z\sum_{i=1}^{l}\log\left(1+e^{-c_i a^T d_i}\right)$$

Where $z>0$ is a parameter chosen by the users.

2.9 Multivariate Regression:

Multivariate regression model is used to solve modern statistical problems. The main task of multivariate regression is to make statistical inference for a possibly sparse and low-rank coefficient matrix C, so the meaningful dependence structure between the responses and predictors can be revealed. In multivariate regression the number of

response variables and the predictor variables has high dimensionality [54].

Let the response vector be $a_i \in R^y$ and the predictor vector $b_i \in R^x$ where $i = 1, 2, ..., N$.

The multivariate regression model can be expressed as,

$$A = BC + E \quad (20)$$

Where $A = (a_1, a_2, ..., a_n)^T \in R^{n \times y}$ is the response matrix $B = (b_1, b_2, ..., b_n)^T \in R^{n \times x}$ is the predictor matrix, $C \in R^{x \times y}$ is the regression coefficient matrix and $E = (e_1, e_2, ..., e_n)^T \in R^{n \times y}$ is the error matrix.The extension of the multivariate regression model can be expressed as

$$A_i = U(B_i C + E_i), \text{where} \, i = 1, 2, ..., N \quad (21)$$

In the above expression $U : R \to R$ is a monotonic function called as utility (or) link function [55].

2.10 Multiple Regression

The multiple regression algorithm is used when more than one regressor variable is occurred. The multiple regression analysis provides an approach to understand the mutual effects of explanatory variable on the response variable.Consider v_m is the response variable for each m, where v_m depends on the regressor variable u_{mn} and the error function ε_m, β_n is the coefficient of the regressor, β_0 is the intercept of the model [56].A multiple regression model is specified as [57],

$$v_m = \beta_0 + \sum_{m=1}^{z} u_{mn} \beta_n + \varepsilon_m \text{ where, } m = 1, 2, ..., N \quad (22)$$

The error term is also called as the residual of the observation. Multiple regression model can be expressed in matrix form as,

$$v = U\beta + \varepsilon \quad (23)$$

Where, $v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_z \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}$

The basic representation of multiple regression for N variables is, $v = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + ... + \beta_N u_N$ (24)

Where β is the regression coefficient.

III. RESULTS AND DISCUSSION

We have presented two different types of data mining algorithms such as classification and regression.. Each type of algorithms has its own strength and weakness. These two types of algorithms are also used in different areas such as, market analysis, fraud detection, production control, customer relationship, etc. So we cannot decide which is the best type of algorithm. But the classification algorithms have more advantages than the regression algorithm. Based on the above comparison, we can say that the classification algorithm is better than the regression algorithm.

Table 1. Comparison of different data mining algorithms and their parameters

Algorithm	Strengths	Limitations	Application

KNN [13]	Simple to understand, Easy to implement, High accuracy, Improved run time.	Poor run time, Sensitive to datas, Outperformed in difficult tasks	Text & Image classification, Intrusion detection, Bio-informatics.
NB [17]	High accuracy, Simple, robust, effective, Reduced computational cost	Data scarcity, Continuous features Makes strong assumptions	Web mining, Fraud detection, Image classification, Text classification
DT [24]	Easy to use, Understandable structure, Human readable rules, Process numerical &categorical variables.	Requires large memory, Need to sort all variables, Needs more time to run.	Fraud detectionHealth care Business & management, Customer relationshipEngineering
NN [32]	High accuracy, Ease of handling, Insensitivity, No manual detection	Longer learning time, Gives low error rate, Difficult to expressclassification rules	Fraud detection, Telecommunication, Medicine, Marketing, Insurance, Finance
SVM [35]	High accuracy, High classification speed, Good generalization ability	Slow training, Complex algorithm, Difficult to implement	Object recognition, Text classification, digit Recognition
Simple Linear Regression [44]	Linearity, Reveal changes in the, covariates time	Impose strong constraint, Add more error bias, Required practicalapplications	Automobiles, Data analysis, Biological system, Astronomy
LASSO Regression [46]	Sensitivity, Specificity, Robustness	Expensive, Need stepwise calculations	Astronomy, Genomics, Feature extraction
Logistic Regression [48]	Allows the modelling of predicted, unbalance &small datasets	Performance is slow, Increased empirical coverageprobability, Increased precision	Document classification, Natural language, Processing
Multivariate Regression [54]	Ability to identifyanomalies, Analysing dataProvides powerful test ofsignificance	Complex, Requires high level, mathematics, More expensive	Scientific research, Genomics, Economics, Image processing, Astronomy
Multiple Regression [55]	High accuracy, Simplicity, SpeedCapacity of learning	Low performance, More expensive	Job design, Leadership, Expectancy theory, Geomechanics

IV. CONCLUSION

Data mining is one of the solutions for the information explosion in current information

institutions. This paper has illustrated the techniques involved in the data mining process, and different types of algorithms used to perform the data mining concept. In this paper we have presented the two types of data mining algorithms (classification & regression), and also the different types of classifiers and regression models present in those types of algorithms. We have also examined the advantages, disadvantages of each type of classifiers and regression models and finally we have provided the comparison for those different types of algorithms.

V. REFERENCES

[1]. Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. "Data mining: an overview from a database perspective." IEEE Transactions on Knowledge and data Engineering vol.8, no.6, pp.866-883, 1996.

[2]. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM vol.39, no.11, pp.27-34, 1996.

[3]. Chauhan, Divya, and VarunJaiswal. "An efficient data mining classification approach for detecting lung cancer disease." Communication and Electronics Systems (ICCES), International Conference on.IEEE, 2016.

[4]. Lakshmi, B. N., and G. H. Raghunandhan. "A conceptual overview of data mining." Innovations in Emerging Technology (NCOIET), 2011 National Conference on.IEEE, 2011.

[5]. Keim, Daniel A. "Information visualization and visual data mining." IEEE transactions on Visualization and Computer Graphics vol.8, no.1, pp.1-8, 2002.

[6]. Tsui, Kwok-Leung, et al. "Data mining methods and applications." Springer handbook of engineering statistics.Springer London, pp.651-669, 2006.

[7]. Jackson, Joyce. "Data mining; a conceptual overview." Communications of the Association for Information Systems, vol.8, no.1, pp.19, 2002.

[8]. Chen, Sherry Y., and Xiaohui Liu. "The contribution of data mining to information science." Journal of Information Science, vol.30, no.6, pp.550-558, 2004.

[9]. Zaki, Mohammed J. "Parallel and distributed data mining: An introduction." Lecture Notes in Computer Science (2000): 1-23.

[10]. Sousa, Tiago, Arlindo Silva, and Ana Neves. "Particle swarm based data mining algorithms for classification tasks." Parallel Computing 30.5 (2004): 767-783.

[11]. Song, Yunsheng, et al. "An efficient instance selection algorithm for k nearest neighbor regression." Neurocomputing, vol.251, pp.26-34, 2017.

[12]. Zhang, Min-Ling, and Zhi-Hua Zhou. "A k-nearest neighbor based algorithm for multi-label classification." Granular Computing, 2005 IEEE International Conference on.Vol. 2.IEEE, 2005.

[13]. Cunningham, Padraig, and Sarah Jane Delany. "k-Nearestneighbour classifiers." Multiple Classifier Systems vol.34, pp.1-17, 2007.

[14]. Alkhatib, Khalid, et al. "Stock price prediction using k-nearest neighbor (knn) algorithm." International Journal of Business, Humanities and Technology vol.3, no.3, pp. 32-44, 2013.

[15]. Adeniyi, D. A., Z. Wei, and Y. Yongquan. "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method." Applied Computing and Informatics vol.12, no.1, pp. 90-108, 2016.

[16]. Imandoust, SadeghBafandeh, and Mohammad Bolandraftar. "Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background." International Journal of Engineering Research and Applications vol.3, no.5, pp.605-610, 2013.

[17]. Choubey, Dilip Kumar, et al. "Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection." Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016). 2017.

[18]. Rennie, Jason D., et al. "Tackling the poor assumptions of naive bayes text classifiers." Proceedings of the 20th International Conference on Machine Learning (ICML-03). 2003.

[19]. McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." AAAI-98 workshop on learning for text categorization.Vol. 752. 1998.

[20]. Arar, ÖmerFaruk, and KürşatAyan. "A Feature Dependent Naive Bayes Approach and Its Application to the Software Defect Prediction Problem." Applied Soft Computing 2017.

[21]. Yang, ChuanChoong, Chit Siang Soh, and VooiVoon Yap. "A non-intrusive appliance load monitoring for efficient energy consumption based on Naive Bayes classifier." Sustainable Computing: Informatics and Systems vol.14, pp.34-42, 2017.

[22]. D'Agostini, Giulio. "A multidimensional unfolding method based on Bayes' theorem." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment vol.362, no.2-3, pp.487-498, 1995.

[23]. Leung, K. Ming. "Naive bayesian classifier." Polytechnic University Department of Computer Science/Finance and Risk Engineering 2007.

[24]. Yu, Zhun, et al. "A decision tree method for building energy demand modeling." Energy and Buildings vol.42, no.10, pp.1637-1646, 2010.

[25]. Jin, Chen, Luo De-Lin, and Mu Fen-Xiang. "An improved ID3 decision tree algorithm." Computer Science & Education, 2009.ICCSE'09.4th International Conference on. IEEE, 2009.

[26]. Liu, Wei, et al. "A robust decision tree algorithm for imbalanced data sets." Proceedings of the 2010 SIAM International Conference on Data Mining.Society for Industrial and Applied Mathematics, 2010.

[27]. Pal, Mahesh, and Paul M. Mather. "An assessment of the effectiveness of decision tree methods for land cover classification." Remote sensing of environment vol.86, no.4, pp. 554-565, 2003.

[28]. Meng, Qi, et al. "A communication-efficient parallel algorithm for decision tree." Advances in Neural Information Processing Systems. 2016

[29]. Li, Ye, et al. "Privacy-Preserving ID3 Data Mining over Encrypted Data in Outsourced Environments with Multiple Keys." Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference on.Vol. 1.IEEE, 2017.

[30]. Devasenapathy, K., and S. Duraisamy. "Evaluating the Performance of Teaching Assistant Using Decision Tree ID3 Algorithm." Entropy vol.151, no.49, pp.0-325, 2017.

[31]. Yang, Shuo, Jing-ZhiGuo, and Jun-Wei Jin. "An improved Id3 algorithm for medical data classification." Computers & Electrical Engineering 2017.

[32]. [32Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining." Advances in Cryptology—CRYPTO 2000. Springer Berlin/Heidelberg, 2000.

[33]. [33Singh, Yashpal, and Alok Singh Chauhan. "NEURAL NETWORKS IN DATA MINING." Journal of Theoretical & Applied Information Technology vol.5, no.1, 2009.

[34]. [34Saxena, Abhinav, and Ashraf Saad. "Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems." Applied Soft Computing vol.7, no.1, pp.441-454, 2007.

[35]. [35Noble, William S. "What is a support vector machine?." Nature biotechnology vol.24, no.12, pp.1565-1567, 2006.

[36]. [36Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning vol.46, no.1,pp.389-422, 2002.

[37]. [37Pal, Mahesh, and P. M. Mather. "Support vector machines for classification in remote sensing." International Journal of Remote Sensing vol.26, no.5, pp.1007-1011, 2005.

[38]. [38Cao, Li-Juan, and Francis Eng Hock Tay. "Support vector machine with adaptive parameters in financial time series forecasting."

IEEE Transactions on neural networks vol.14, no.6,pp.1506-1518, 2003.

[39]. [39Zeng, Zhi-Qiang, et al. "Fast training Support Vector Machines using parallel sequential minimal optimization." Intelligent System and Knowledge Engineering, 2008.ISKE 2008.3rd International Conference on. Vol. 1.IEEE, 2008.

[40]. [40Widodo, Achmad, and Bo-Suk Yang. "Support vector machine in machine condition monitoring and fault diagnosis." Mechanical systems and signal processing, vol.21, no.6, pp.2560-2574, 2007.

[41]. [41Chen, Ting, AnandRangarajan, and Baba C. Vemuri. "Caviar: Classification via aggregated regression and its application in classifying oasis brain database." Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on.IEEE, 2010.

[42]. [42Zou, Kelly H., Kemal Tuncali, and Stuart G. Silverman. "Correlation and simple linear regression." Radiology vol.227, no.3,pp.617-628, 2003.

[43]. [43Naseem, Imran, Roberto Togneri, and Mohammed Bennamoun. "Linear regression for face recognition." IEEE transactions on pattern analysis and machine intelligencevol.32, no.11,pp.2106-2112, 2010.

[44]. [44Aalen, Odd O. "A linear regression model for the analysis of life times." Statistics in medicine vol.8, no.8,pp.907-925, 1989.

[45]. [45Chang, Le, Steven Roberts, and Alan Welsh. "Robust Lasso Regression Using Tukey'sBiweight Criterion." Technometricsjust-accepted, 2017.

[46]. [46Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological),pp.267-288, 1996.

[47]. [47Lu, Yiming, et al. "A Lasso regression model for the construction of microRNA-target regulatory networks." Bioinformatics vol.27, no.17,pp.2406-2413, 2011.

[48]. [48Hosmer, David W., et al. "A comparison of goodness-of-fit tests for the logistic regression model." Statistics in medicinevol.16, no.9, pp.965-980, 1997.

[49]. [49Dreiseitl, Stephan, and LucilaOhno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics vol.35, no.5, pp.352-359, 2002.

[50]. [50Hilbe, Joseph M. "Logistic regression." International Encyclopedia of Statistical Science. Springer Berlin Heidelberg, pp.755-758, 2011.

[51]. [51Cepeda, M. Soledad, et al. "Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders." American journal of epidemiology vol.158, no.3, pp.280-287, 2003.

[52]. [52Goh, Gyuhyeong, Dipak K. Dey, and Kun Chen. "Bayesian sparse reduced rank multivariate regression." Journal of Multivariate Analysis vol.157,pp.14-28, 2017.

[53]. [53Kharratzadeh, Milad, and Mark Coates. "Semi-parametric order-based generalized multivariate regression." Journal of Multivariate Analysis vol.156,pp.89-102, 2017.

[54]. [54Jönsson, Carl Axel, and Emil Tarukoski. "How does an appointed ceo influence the stock price?: A Multiple Regression Approach." 2017.

[55]. [55Singh, Rajesh, et al. "Prediction of geomechanical parameters using soft computing and multiple regression approach." Measurement vol.99, pp.108-119, 2017.