REST Journal on Emerging trends in Modelling and Manufacturing
Vol:4(4),2018
REST Publisher
ISSN: 2455-4537

Website: www.restpublisher.com/journals/jemm

Incremental Knowledge Mining Process Based on Supervised and Unsupervised Learning Dataset

Niranjan Shrivastava[1], Deepak Sukheja[2], Prateek Sharma[3]
[1]Associate Professor, IMS, DAVV, Indore
[2]Associate Professor, VNR, VJIETT, Hyderabad.
[3]Associate Professor, PIMR, Indore.
nirshri@gmail.com

Abstract

The process of mining to information and knowledge from the huge data has been authored by several researchers as a core research area in database systems, Data warehouse and mining, Big Data and machine learning. Also, the process of information and knowledge mining has been used by different types of organizations with an opportunity to generate the good revenues and expansions of their business by predicting the future scenario. In present scenario knowledge mining process is a subset of machine learning, Artificial Neural Network (ANN) and Artificial Intelligence (AI). This paper presents an incremental knowledge mining process using the documental revision techniques of data, data collection methods and data mining techniques.

Keywords: *Data, Data collection techniques, Data mining techniques and knowledge mining process.*

I. Introduction

Data, Data Collection and Data Mining are most common and integrated terms in analytical research, Machine Learning and research in decision making process. This is since data plays a pivotal role among all latest technologies like DBMS, RDBMS, DDBMS/HDBMS, DW(Data Warehousing), Cloud Computing, Big Data and Big Data Analytics. All these tools and techniques are used to provide the information for the transaction system or analytical processes. They differ on the Size of database can range from Megabyte to Terabyte, Data retrieval process and Data Visualization Process. The data attributes viz. type of data, quality of data and consistency of data are common among these technologies. But the consistency of data is dependent on data collection techniques. Similarly, the output and visualization of technologies are dependent on data mining techniques. The prerequisite of data mining technique is a well-structured dataset in any format. The data inside the dataset may be unlabelled or labelled. Therefore, this paper is categorised into four sections, there are Data, Data Collection Techniques, Data identification and separation and Knowledge Mining Process. First section of this paper describes in general about the data, its definition and related information concerning it. Second section is illustrating important data collection techniques. The third section exemplifies different data mining techniques for segregate the collected data into unsupervised or supervised dataset and fourth section defines the knowledge mining process.

1. Data

As a matter of fact, the term data conveys the meaning of a simple entity. But realistically it is very difficult and complex to define the term data. Keeping in view associated complexities, many definitions are coined to define the term data as Data constitutes distinct pieces of information, usually formatted in a special way. Data could be referred to as an entity that is recorded and stored. In terms of computer system, data is a Fact, text, graphics, image, audio and video that has meaning in the user's environment. In terms of mathematics, data is only a figure which is required to generate a mathematical model or generated through mathematical model. According to [1], the definition of data in research perspective is sub divided into two levels; Primary Data and Secondary Data. Primary data inculcate the data originated by the researcher for the first time through direct efforts and experience, specifically oriented towards addressing the concerned research problem. This is also known as the first hand or raw data. Secondary data gives an indication of second-hand information which is already collected and recorded by any person other than the user for a purpose, not relating to the currently addressed research problem. In other words, it is generally concluded that Primary data is a real-time data whereas secondary data is the one which relates to the past. Based on characteristics of the data, it could be categorised as qualitative or quantitative, it may be continuous or discrete, it may be open or attribute or it may be nominal or ordinal. The hierarchy of data is shown in figure 1.

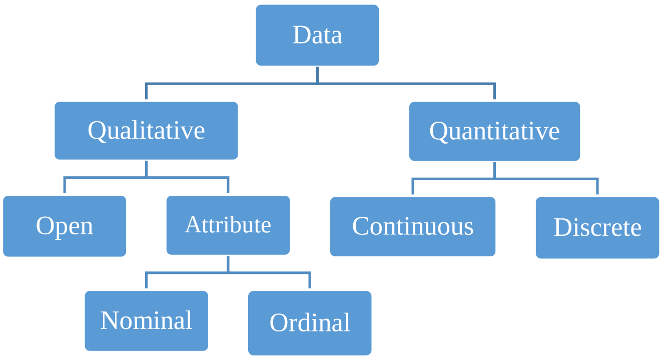

Figure 1: Hierarchy of data

1.1 Qualitative Vs Quantitative

As defined in [1], Qualitative Data belongs to the category of data that provides insights and understanding about a problem. The speciality of qualitative data is that it can be approximated but cannot be computed. Hence, it should be imperative to the researcher that complete knowledge about the type of characteristic needs to be encompassed, prior to the collection of data i.e. Qualitative data cannot be expressed as a number (The nature of data is descriptive and so it is a bit difficult to analyse it). Data that represent nominal scales such as gender, social economic status and religious preference are usually considered to be qualitative data. Exploration of research methodology is perceived using qualitative data. It helps to build understanding and provides insight. Quantitative Data, as per the name deals in quantity or numbers. It is an indicative of the data that is used for computation of the values, counting and can be expressed in numerical terms. The most prominent deployment of this category of data is in statistical information retrieval and manoeuvring. This type of data is conclusive in nature which is targeted at testing a specific hypothesis and examines the relationships.

1.1.1 Continuous Vs Discrete Data

Previous section describes two important categories of data, qualitative or quantitative. Quantitatively it is again bifurcated into continuous and discrete. Continuous data represents collection of information that can be measured on a continuum or scale. It encompasses almost any numeric value. It could also be subdivided into finer and finer increments, based on the precision of the measurement system [2]. At other end, discrete data stands for unified data that cannot be sub divided further; it is distinct and can only occur in certain values. Conclusively it is indicated that discrete data is counted, Continuous data is measured.

1.1.2 Open Vs Attribute Data

Open data belongs to the category of data that can be freely used, re-used and redistributed by anyone i.e. The data must be provided under terms that permits re-use and redistribution including the intermixing with other datasets.

1.1.2.1 Nominal Vs ordinal data

Nominal values are observations that can be assigned a code in the form of a number, where the numbers are simply labels. On the other hand, ordinal values or observation (put in order) have a rating scale attached. Ordinal data could be ordered and counted but could not be measured.

II. Data Collection Techniques

Data collection be a main functional component of research in all fields of study not only in computer science. It also includes literatures, social sciences, humanities and business. While the data collection methods may vary by disciplines. Data collection is a concept of collecting information to address the critical evaluation questions that has been identified earlier in the evaluation process. As earlier described in section 1, there are two main types of data that users find themselves working with – and having to collect Quantitative and Qualitative data, therefore classification or categorization of data collection methods again is based on qualitative data or quantitative data.

A. Qualitative Data Collection Methods

Generally, qualitative data collection methods are time-consuming and expensive to collect the data. The common and most reliable qualitative data collection methods are:

- Face-to-Face Personal Interviews
- Qualitative Surveys
- Web-based questionnaires
- Focus Groups
- Documental Revision
- Observation
- Case Studies

i. Face to Face Personal Interviews

Face to Face Personal Interview technique is most common data collection mechanismto collect qualitative data for research purpose because of its personal approach. The interviewer will collect data directly from the target/subject/object (the interviewee), on a one-on-one and face-to-face interaction. This is ideal when data to be obtained must be highly personalized. The data collected through this technique may be informal, unstructured conversational. In the case of planned interview, the collected data may be semi-structured.

ii. Qualitative Surveys

Qualitative Survey technique is a very common and effective technique to collect the semi-structured and structured data. The concept used in this technique is Paper surveys, questionnaires and web-based questionnaires etc. The quality and reliability of the data is also depending on the prepared questionnaires, type of surveys and

questionnaires is very popular technique to collect structured data. In this technique,questionnaire is uploaded to a site, where the respondents will log into and accomplish electronically (some time all possible answer also enclosed with questions).

iii. Focus Groups

The focus group data collection technique is same as a face to face interview, the difference is the interview has been conducted in the form of group discussions. To collect the versatile qualitative research type of data focus group approach is highly recommended. Most recent examples of this technique is to collect the data to find out the impact of demonetization of Indian currency or impact of GST in middle level industrial sector. The data collection is not possible through only personal interview, questionnaires therefore the multiple group discussion (with experts)has been conducted to find the views and collect data on these topics.

iv. Documental Revision

As defined in [3], Documental revision technique makes use of existing and reliable documents along with other reliable sources of information as a source of data. It isutilized in a new research or investigation viz. literature review or literature survey. This is relevant to the process followed by the data collector, going to the library, go over the books and other relevant documents corresponding to his recent research.

v. Observation

Observation is a most reliable data collection technique but it is very sensitive and specific relevant to the topic. This technique relies on the measurement procedure or observed behaviour with respect to provide input of an instrument/ experiment. The basic disadvantage of this technique is; some time the cost of data is very high and/or some time it is very time consuming for the users to collect the data because to obtain reliability, behaviours must be observed several times. Data can be interpreted differently using the following mechanisms:

1. Descriptive observations: Jotting down the observations straight way from the observations.
2. Inferential observations: Writing down the observation indirectly e .g. physical behaviour of the subject.
3. Evaluative observation: Assuming and hence be judgmental from the behaviour. Also, ensuring the replication of these findings.

vi. Case Studies / Document Review

A case study is depicted as an in-depth description of a process, experience, or structure focussed on a single institution. The ‘what’ and ‘why’ questions are answered using the case studies, which involves a mix of quantitative (i.e., surveys, usage statistics, etc.) and qualitative (i.e., interviews, focus groups, extant document analysis, etc.) data collection techniques [4].

Intuitively, the quantitative data is analysed at first and then qualitative strategies are utilised to investigate the trends identified in the numerical data.

B. Quantitative Data Collection Methods

Common and most reliable quantitative data collection methods are:

- Quantitative Surveys
- Computer-assisted interviews
- Quantitative Observation
- Experiments: Laboratory, Field, Natural Experiments.

i. Quantitative Surveys

As mentioned in sub section 2.1.2, unlike the open-ended questions asked in qualitative questionnaires, quantitative paper takes care of closed questions, supplemented with the answer options provided. The respondents answer among the choices provided on the questionnaire. This is very common method to get used by researcher to define the recent trends and mentality about current scenario.

ii. Computer-assisted interviews

This is called CAPI, or Computer-Assisted Personal Interviewing which consists of face-to-face interviews, the data obtained from the interviewee will be entered directly into a database using a computer.

iii. Quantitative Observation

This is a form on which observations of an object or a phenomenon are recorded. The items to be observed are determined regarding the nature and objectives of the study. They are grouped into appropriate categories and listed in the schedule in the order in which the observer would observe them.

The schedule must be as devised as to provide the required verifiable and quantifiable data and to avoid selective bias and misinterpretation of observed items. The units of observation must be simple, and meticulously worded to facilitate precise and uniform recording.

iv. Experiments

Quantitative researches often make use of experiments to gather data, and the types of experiments are:

Laboratory experiments: This is your typical scientific experiment setup, taking place within a confined, closed and controlled environment (the laboratory), with the data collector being able to have strict control over all the variables. This level of control also implies that he can fully and deliberately manipulate the independent variable. Field experiments: This takes place in a natural environment, "on field" where, although the data collector may not be in full control of the variables, he is still able to do so up to a certain extent. Manipulation is still possible, although not as deliberate as in a laboratory setting. Natural experiments: This time, the data collector has no control over the independent variable whatsoever, which means it cannot be manipulated. Therefore, what can only be done is to gather data by letting the independent variable occur naturally, and observe its effects for example find the pollution level in air through MQ 135 sensor, this sensor identifies the level of $PM_{2.5}$, $PM_{2.5}$ is independent variable for the sensor. Sensor will check and record $PM_{2.5}$ level in air. As per the above discussion there are different data collection techniques available. The basic function of data collection technique is collection of primary or secondary data in terms of the quantitative or qualitative so that it can be bifurcate according to the figure 1. The mandatory requirements at the time of data collections are be clear of your purpose, Define the scope of study, Develop a research question and develop a list of research objectives. With the consideration of above requirement researcher design the structure of the dataset in terms of attribute in which labelled or unlabelled data could be store. The collection of the data in attribute may be in different file like CSV, DBF, XLS, ARRF etc but the basic objective of dataset is to work as an input for machine learning or artificial neural network algorithms to produce accurate result.

III. Data Identification and Separation

The data identification and separation concept is very important and ply very crucial role in data mining, machine learning algorithm or to design artificial neural network model. On behave of separation of labelled or unlabelled data from collected dataset, different data mining techniques can be applied. Normally we apply unsupervised learning algorithm for unlabelled data and supervised learning algorithm for labelled data. The basic difference between unlabelled and labelled data is, unlabelled data consists of samples of natural or human-created artefacts that you can obtain relatively easily from the world. There is no "explanation" for each piece of unlabelled data "it just contains the data and nothing else", and Labelled data typically takes a set of unlabelled data and augments each piece of that unlabelled data with some sort of meaningful "tag," "label," or "class" that is somehow informative or desirable to know. There are many scenarios where unlabelled data is teeming and easily obtained but labelled data often requires a expertization or mathematically formulated to interpret.For example, suppose one agency want to find the employed or unemployed peoples of a country, the task is very simple but when agency start the survey and to collect the data related with their income, it will very difficult to segregate the people between employment and unemployment. Because of employment may be categories at number of vertical level in which some are labelled and many are unlabelled. In context to our country (India) As per the definition of labelled data, a personwho is paying a tax he/she is come under the employed categories,because of taxation is compulsory for salaried or self-employed person if earning is greater than a fixed amount(like in India 2.5 lake per year, formal jobs or business in formal sector/ as per government rules). At the same time,huge self-employed and salaried employee is earning less than 2.5 lake per year (working in informal sector) and number of person or self-employed those are working in informal sector and earning more the 2.5 lake per year and not paying tax i.e. unlabelled data.Therefore, the identification and separation of labelled and unlabelled data is very important.

A. Data Pre-processing

Data pre-processing is the substantial step in data mining. Data mining mainly involve missing value replacement, transformation, normalization, and discretization. The result of data pre-processing is the final training set.

i. Missing value replacement

There are many originators of missing value such as broken sensor, erroneous or missing data entries and in some case some attributes make no sense for some type of objects. It is necessary to replace the missing value otherwise the analysis could lead to meaningless denouement. The clear way is to replace the missing value by mean value in case of numeric attributes and modus in case of categorical value.

ii. Normalization

Normalization is also important task in data pre-processing to reduce unwanted variation either within or between arrays. Normally normalization can be done on data with three ways such as Z-Score, by decimal scaling or min-max normalization. There are main two types of normalization based on a) distance and b) proportion. Distance based normalization includes vector based that is on Euclidian distance and linear based normalization which can correct skewedness in data. Proportion normalization includes nonmonotonic normalization which is on Z-Score. The normalization property requires that the range of a sameness or distance measure lies within a fixed range.

iii. Transformation

Transformation is also a valuable step in data pre-processing. Transformation almost compresses the maximum data. Transformation mainly involves smoothing, aggregation, generalization and discretization. A. Kusiak et al. introduced new transformation method named feature bundling [5]. When this transformation technique applied to a training data set it embellish classification accuracy of the decision rules generated from this set. Although bundling is destined for integer, categorical and normative features, it can be continued with continues value, for example by using regression function.

iv. Discretization

Discretization is a process of transforming continuous attribute value into finite set of intervals to generate attributes with a smaller number of distinct values. There are many types of discretization methods such as Direct vs. Incremental, Single vs. Multi attribute, supervised vs. Unsupervised, Bottom up vs. Top down.

B. Data Exploration

Data Exploration is about describing the data by means of statistical and visualization techniques. We explore data to bring important aspects of that data into focus for further analysis. Data exploration is of following two types.

1. Univariate analysis
2. Bivariate analysis

i. Univariate Analysis

The variables are explored one by one during univariate analysis it may be continuous or categorical based on the type of the variables used.

a. Continuous Variables: The central tendency and spread of the variable represents its nature.

b. Categorical Variables: The distribution of each variable is represented by frequency table. The percentage of values under each category is measured using two metrics namely count and count%.

ii. Bivariate Analysis

It shows the relationship between two variables. This also reveals the association and disassociation between variables at a pre-defined level. This may be viewed as a combination of categorical &categorical, continuous& continuous and categorical & continuous.

a. Continuous & Continuous:

The relation between two continues variables may be linear or non-linear. Bivariate analysis only shows the relationship between two variables not the strength of relationship among them. This can be done by evaluating correlation between the variables and can be expressed as:

$$Correlation = Cov(X,Y) / \sqrt{(Var(X) * Var(Y))} \quad (1)$$

The range of correlation always ranges between [-1,1].

b. Categorical & Categorical

The relation between two categorical variables may be analysed either by a two-way table, stacked column chart and by chi-square test.

c. Categorical & Continuous

The relation between categorical and continuous variables will not show the statistical significance if the levels for each category is small. It can be done by performing Z-test and ANOVA.

IV. DATA MINING TECHNIQUES

Data Mining is an analytic process designed to explore data i.e. data mining techniques are to extract the required data or select specific data related with the task then design and visualize the patterns for the decision support system. Various data mining techniques are available to implement decision support system. These data mining techniques are described at two levels: predictive and descriptive. Predictive technique and descriptive techniques also categorized at different following levels.

1. Predictive Techniques:
   - Classification
   - Regression
   - Time series analysis
   - Prediction
2. Descriptive techniques:
   - Clustering
   - Summarization
   - Association rules
   - Sequential discovery

A. Predictive data mining Technique

The goal of data mining is prediction. Predictive data mining is the most common type of data mining which has the most direct business applications.

i. Classification

Classificationtechnique consists of predicting a certain outcome based on a given input. Classification techniques in data mining can process a large amount of data. It can be used to predict categorical class labels, classifies data based on training set and class labels and it can be used for classifying newly available data. This technique has been implemented in two steps; During first step the model is created by applying classification algorithm on training data set then in second step the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So, classification is the process to assign class label from dataset whose class label is unknown.

a. K-Nearest Neighbour

K-Nearest Neighbour (KNN) is the type of supervised learning method. The working process of KNN classifier is defined below:

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 neighbours (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,

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

$$D(p,q)=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad (3)$$

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 [6]. Let C be the class membership of KNN classification. The membership of test data set can be calculated by using the following expression.

$$C_i=\{p \in C_n ; D(p,p_i) \leq D(p,p_r), i \# r\} \quad (4)$$

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

b. Naïve Bayes

The Bayesian classifier is constructed based on the Bayesian network [7]. Naive Bayes classifier is a linear classifier, in which the attributes are considered as independent and have equal weight. 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 [8]. It can be expressed as,

$$P(H/E)=\frac{P(E/H)\ P(H)}{P(E)} \quad (5)$$

c. Decision Tree

Decision tree (DT) 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 [9]. 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). $BestSpli$returns the best split point and $FindSpli$ splits the data according to the $BestSpli$ point [10].

d. Neural Network

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) [11]. 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 [12].

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

e. Support Vector Machine

Support Vector Machine (SVM) is a computer algorithm which learns by example to assign labels to objects. Consider a problem of classifying m points into n-dimensional real space R^N which can be represented as $m \times n$ matrix [13]. 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 [14]. The classification hyper plane in N - dimensional space is $\omega\ a+z=0$. Consider a hyper plane $f(X)=0$,which separates the data. $f(X)=\omega^T a+z=\sum_{y=1}^{N}\omega_y a_y+z=0$ (6)

where ω is a vector on N dimensional space and z is a scalar. $b_x\ f(X_x)=b_x(\omega^T a_x+z)\geq 1$ (7)

where $x=1,2,...,N$. The QP problem is expressed as [15], $\min \phi(\omega)=\frac{1}{2}\|\omega\|^{2}$ (8)

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

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

$$\min \phi(\omega,\xi)=\frac{1}{2}\|\omega\|^{2}+C.\left(\sum_{x=1}^{N}\xi_x\right) \quad (9)$$

Where $b_x\left[\omega.a_x+z\right]\geq 1-\xi$, $x=1,2,3,....N$, $\xi_x\geq 0$, $x=1,2,3,....N$.

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

$$\max W(\alpha)=\sum_{x=1}^{N}\alpha_x-\frac{1}{2}\sum_{x,y=1}^{N}\alpha_x\alpha_y b_x b_y\left(a_x,a_y\right) \qquad \text{Where } 0\leq\alpha_x\leq C,\quad x=1,2,...,N \quad (10)$$

$$\sum_{x=1}^{N}b_x\alpha_x=0 \quad (11)$$

By solving the above problem, the classifier can be expressed as, $f(X)=sign\left(\sum_{x=1}^{N}\alpha_x b_x\left(a\ .\ a_x\right)+y\right)$ (12)

Where α_x is the solution of QP problem [16].

ii. Regression

A Regression is a data mining technique which forecasts a range of numeric values (referred as *continuous values*), for a given dataset. Regression is a data mining (machine learning) technique which fits an equation to a dataset. Regression analysis is a form of predictive modelling technique which considers the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for prediction. Example: regression is 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 modelling and analysis of trends. There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).

a. Linear Regression

It is considered as one of the key modelling technique. In this technique, continuous attribute is asserted to the dependent variable, while independent variable(s) can be continuous or discrete. The regression line is linear. It establishes a relationship between dependent variable (Y) and one or more independent variables (X)by considering a best fit straight linei.e regression line. The relation can be expressed as,

$$Y_i=a+b\,X_i+e_i \quad (13)$$

where a is intercept, b is the slope of the regression line, e_i is the random error term and $i=1,2,...,N$.

b. Logistic Regression:

Logistic regression is used to find the probability of success event and failure event. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. The probability can be calculated as,

$$P\left(c=\pm 1|d,a\right)=\frac{1}{1+\exp\left(-c\left(a^{T}d+b\right)\right)} \quad (14)$$

where data d , weights (a,b) and class label c , d_i is a training instance, $i=1,2,...,l$ and $c_i\in\{1,-1\}$.

c. Polynomial Regression:

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

d. Stepwise regression:

This regression type is used to deal with multiple independent variables. This technique performs the selection of independent variables using an automatic process, and this involves no human intervention. This is done by noting statistical values like R-square, t-stats and AIC metric making use of significant variables. It best fits the regression model by adding/dropping co-variants one at a time based on a specified criterion.

iii.Time series analysis

Time series analysis is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is a method of using a model to generate predictions (forecasts) for future events based on known past events. For example, stock market.

iv.Prediction

It is one of a data mining techniques that discover the relationship between independent variables and the relationship between dependent and independent variables. Prediction model is based on continuous or ordered value.

B. *Descriptive techniques*

Another approach of data mining to extract the knowledge from large datasets is known as Descriptive data mining. It is normally used to generate correlation, frequency, cross tabulation, etc. This Descriptive method can be defined as to discover regularities in the data and to uncover patterns.

i. Clustering

Clustering is a separation of data into groups of similar objects, disparate object into another cluster. It is a way of finding similarities between data according to their quality. This technique based on the unsupervised learning. It is also categorized at following different level;

Partitional Clustering: A splitting upof data objects into non-overlapping subsets (clusters) such that each data object is subset.

Hierarchical clustering: A set of nested clusters organized as a hierarchical tree.

The different Clustering Algorithms areK-Means, single linkage algorithms and simulated annealing (SA) based clustering technique

ii. Summarization

Summarization is the process of reducing the huge volumes of data in a meaningful and intelligent fashion with important and relevant features. Summarization techniques like tabulation of the mean and the standard deviations are often implied to analyses and visualize the data and to generate the report automatically.

iii. Association rules

Association rule mining unwraps the pattern that occurs frequently among the data set. It focuses in extracting associations, correlations, frequent sequence, frequent item set and frequent patterns with interestingness among the data set in the data repositories. The association can be expressed as $X \rightarrow Y$.

a. Support:

Support determines how often a rule is applicable to a given data set, i.e., the support of the rule is the percentage of transactions that contain both X and Y among all transactions in the input data set. Support can be computed as probability of the union set X & Y. $Support\left(X \rightarrow Y\right) = P\left(X \cup Y\right) = \frac{n\left(X \cup Y\right)}{N}$ (15)

b. Confidence:

Confidence determines how frequently the items in Y appear in transactions that contain X, i.e., the confidence of the rule is the conditional probability of transactions that contain Y among transactions that contain X. The conditional probability also can be computed through proportion of support. $Confidence\left(X \rightarrow Y\right) = P\left(X / Y\right) = \frac{n\left(X \cup Y\right)}{n\left(X\right)}$ (16)

V. INCREMENTAL KNOWLEDGE MINING PROCESS

After the carefully documental revision of all types of data, data collection and data mining techniques which are mentioned in the previous sections of this paper. There are various ways to find the knowledge from the collected data. Knowledge may be in terms of text, pattern, trends etc. This depends on the combination of data, data collection technique and data mining technique/ machine learning algorithms. The proposed simple incremental knowledge mining process is defined in figure 2.

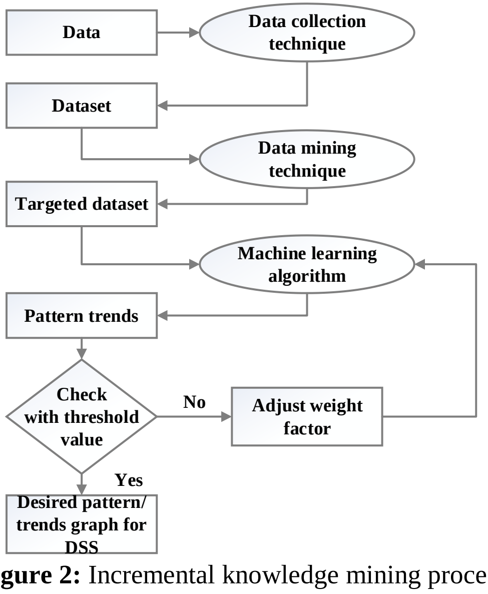
Data
Data collection technique
Dataset
Data mining technique
Targeted dataset
Machine learning algorithm
Pattern trends
Check with threshold value
No
Adjust weight factor
Yes
Desired pattern/ trends graph for DSS

Figure 2: Incremental knowledge mining process

The mentioned process will work as follows:

1. All kinds of data are an input to data collection technique. As per the type of data, structure of data and its characteristics, data collection technique will have used to collect the data. Finally, the data collection technique collects the

data in specific format and returns a data set which holds the labelled and/or unlabelled data with different characteristics in single format.

2. With the help of expertization and mathematical formulation, data identification and separation will have applied to differentiate the labelled and unlabelled data construct two datasets one for labelled data and second for unlabelled data.

3.The different pre-processing technique can apply on labelled and unlabelled dataset to prepared targeted dataset. On huge dataset, which is returned by data collection technique and returned targeted data set after filter to given huge dataset.

4. This specified targeted dataset transfer to the machine learning algorithm or machine learning process which provides the trends and patterns.

5. The range of trends and patterns are normalised between [0-1] by:

$$N(pattern) = \frac{\max(pattern) - current(pattern)}{\max(pattern)} \quad (17)$$

Where $\max(pattern)$ is the maximum pattern value, $N(pattern)$ is the normalised pattern value and $current(pattern)$ is the present pattern value.

6. This trends and patterns will be compared with the threshold value δ, if threshold value will meet then pattern or trends will be accepted by decision support system else the value of targeted dataset transferred to the machine learning algorithm / machine learning process through supervised or unsupervised learning law after applying weight adjustment factors. which can be expressed as:

$$Output = \begin{cases} Accept & , if \;\; pattern \geq \delta \\ Weight\, adjustment & , otherwise \end{cases} \quad (18)$$

7. The process will also terminate if the trends and patterns are same in last three iterations and not to meet threshold values. This kind of patterns and trends are rejected by the designed mechanism and provided the feedback to administration regarding the targeted dataset.

The accuracy of predicted outcomes evaluated for different machine learning algorithms using both qualitative and quantitative collection techniques are presented in table 1.

TABLE 1: Prediction Accuracy of Data Collection techniques

Data collection techniques	KNN	NB	DT	SVM	NN
Qualitative	88	76	86	90	92
Quantitative	86	79	89	91	90

From table 1 it is observed that the prediction accuracy evaluated by the machine learning algorithms shows small variations for both data collection techniques. KNN outputs 88% accuracy with qualitative approach and 86% with quantitative approach, NB outputs 76% accuracy with qualitative approach and 79% with quantitative approach, DT outputs 86% accuracy with qualitative approach and 89% with quantitative approach, SVM outputs 90% accuracy with qualitative approach and 91% with quantitative approach and NN outputs 92% accuracy with qualitative approach and 90% with quantitative approach. However, the prediction accuracy of NN outperforms the other algorithms. The prediction accuracy of the proposed method also depends on the threshold value. Figure 3 shows the prediction accuracy for machine learning algorithms with different threshold values.

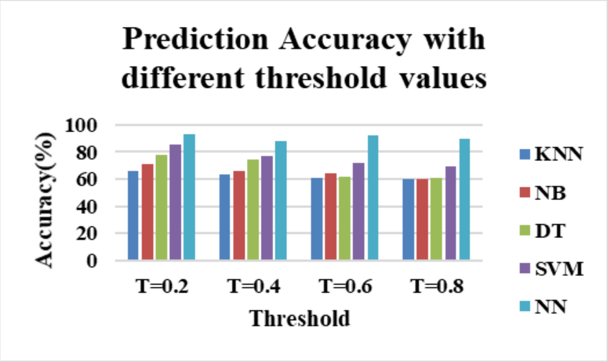

Figure 3: Prediction Accuracy with different Threshold Values

The threshold value is set to 0.2, 0.4, 0.6 and 0.8 and the prediction accuracy for different machine learning algorithms are evaluated and from figure 3 it is observed that if the threshold value increases the acquired pattern/trends decreases. Also, it is visible that NN performs well in all scenario than the other algorithms.

VI. Conclusion

In this paper, we have presented an Incremental Knowledge Mining processusing documental revision techniques for supervised and unsupervised datasets. We have utilized the advantages of several data collection techniques in this paper. The utilization data pre-processing and data exploration also improves the significance of the proposed work. The outcomes of the proposed work presented in this paper is also promising.

Reference

[1]. http://keydifferences.com/difference-between-qualitative-and-quantitative-data.html

[2] www.isixsigma.com/dictionary/continuous-data/

[3] https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

[4].https://alaworkshopdata.wordpress.com/data-collection-tools/

[5] A. Kusiak, Member, IEEE, ―Feature Transformation Methods in Data Mining‖, IEEE transactions on electronics packaging manufacturing, vol. 24, no. 3, July 2001.

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] Saxena, 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.

[13] Pal, 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.

[14] Cao, 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.

[15] Zeng, 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.

[16] Widodo, 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.

[17]. https://alaworkshopdata.wordpress.com/data-collection-tools/

[18]. https://www.scribd.com/document/273130375/Data-Collection

[19]. https://www.thoughtco.com/regression-1019655

[20]. Saja H. Rasool1 , Faiq M.S.Al-Zwainy1 “Estimating Productivity of Brickwork item using Logistic and Multiple Regression Approaches” SJET, 2016; 4(5):234-243

[21] Nikam, Orient. “A Comparative Study of Classification Techniques in Data Mining Algorithms" J. Comp. Sci. & Technol., Vol. 8(1), 13-19 (2015)