IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-4 E-ISSN: 2347-2693

Review Paper on Data Mining and its Techniques and Mahatma Gandhi National Rural Employment Guarantee Act

Kritika Yadav [1*], Mahesh Parmar[2]

[1*]CSE/IT Department, MITS, Gwalior, India
[2] CSE/IT Department, MITS, Gwalior, India
[*]*Corresponding Author: kritika_yadav_13@yahoo.com, Tel.: +91-8004965805*

www.ijcseonline.org

Received: 28/Mar/2017, Revised: 06/Apr/2017, Accepted: 20/Apr/2017, Published: 30/Apr/2017

***Abstract*—** Data Mining is a technique that attempts to find useful pattern from substantial volume of data. The paper reviews data mining and its techniques in e- governance. The paper also reviews the influence of the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) on the rural India. The objective of MGNREGA is to provide at least hundred days of job to the rural and tribal population, whose living entirely depends on daily wages. Moreover the paper gives the relative evaluation of numerous data mining methods and algorithms.

***Keywords*—**Data Mining, MGNREGA, Data Mining Techniques

I. INTRODUCTION

The rising technology usage resulted into generation of enormous quantity of digital data which thereby resulted in large storage database. This expansion of database occurred in many prominent areas like government datum, transaction detail of supermarket, mobile phone call details, and record of credit card usage and also in intricate areas like astronomical data records, medical reports and likes. With the expeditious increase in data, it is a sharp need to extract serviceable information from the database which might result into some advantageous information to the user. This task of exploring data and translating into more meaningful patterns/information is known as data mining[1].

Data Mining is a mechanism of excerpting or mining the knowledge from substantial proportion of data. The expression data mining could be appropriately called as "Knowledge mining". Data congregation and storage technique has made achievable for corporate industry to assemble humongous proportion of data at reduced price. Utilizing this preserved data, so as to infer useful and actionable information, is the universal objective of the common activity termed as data mining. Data mining can be expounded in the following manner:

Data mining is explicated as the procedure of exploring and analysing, by automatic or semiautomatic means, enormous quantity of data in conducive to extract significant rules and patterns. This is a multidisciplinary subfield of information technology that comprise of computation of pattern recognition of big data sets. The objective of this advanced analysis is to deduce knowledge from data set and reorganize it into an intelligible structure for further use. The techniques that are employed are at the confluence of artificial intelligence, machine learning, statistics, database systems and business intelligence. Data Mining is about interpreting problems by examining data already available in databases[2].

Data mining also recognised as knowledge discovery in databases (KDD), where KDD is a conventional method of transfiguring enormous data to consequential interpretation and analysis.

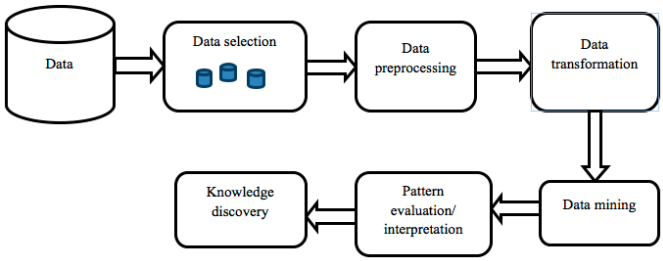


Fig.1 Knowledge discovery process in database

Data mining features are used to state the type of patterns that are developed in data mining procedure. Data mining tasks may be classified into two groups-descriptive and predictive. Descriptive data mining summarize the general characteristics of the data in databases. Predictive data mining task deduce inferences from the existent data for forecasting[2].

The goal of data mining is either to generate a descriptive model or a predictive model. A **descriptive model** depicts

the main features of the data sets. It is substantially the collection of data points, thus accrediting to examine the important facet of data set. Generally, descriptive model is erected over undirected data mining; i.e. a bottom-up approach where the data is "self evident". Undirected data mining detects appropriate patterns from the data set but hand over the perception of the model on the data miner. The purpose of the **predictive model** is to authorize the data miner to speculate an unspecified value of a definite variable i.e., the target variable. If the target value is among predefined number of discrete or class labels, data mining operation is classification. If the required outcome variable is a real number, the task is termed as regression.

Data mining can be grouped into various task based on various objective function. It may be categorized as follows:

- **Exploratory Data Analysis (EDA):** The main objective of EDA is to explore data without having any understanding in advance of what to search in data. This method is very interactive and visual graphical method. EDA works on relatively small dimensional data.
- **Predictive modelling:** In predictive modelling, a model is constructed that foresee the value of dependent variable from the value of independent variables. Distinguished approaches are Classification and Regression. When dependent variable to be predicted is categorical in nature, the technique to be used is classification, while for continuous quantitative dependent variable, regression is used.
- **Descriptive modelling:** Descriptive modelling describes the entire data by generating some information. Furthermore the task is divided including density estimation which gives complete probability distribution of data. Prominent techniques are Clustering and segmentation which are dependent on grouping the data in given dimension space. Dependency modelling is also descriptive modelling that describes relationships between different set of variables.
- **Pattern discovery and rules generation:** It detects meaningful pattern from the database to fulfil any known or unknown objective. For example, the detection of frequent transaction behaviour in any business or detection in space may be grouped under pattern discovery. These classes of patterns can be detected by an algorithmic technique known as association rules.

II. MGNREGA

The Government of India has instigated several employment production programmes to obliterate poverty and unemployment, since 1980. All these programmes were unworthy and gradual in their approach. Therefore, the programmes failed to make any significant impression on the issues of poverty and unemployment. With proliferation and liberation of the economy, it is always distrust that the prevalence of poverty and unemployment will intensify appreciably. In this context, the accomplishment of National Rural Employment Guarantee Act by UPA government is the most appropriate course of action. This iconic programme of UPA government is comprehensive in its aptitude of overall growth and Right to Work. The act was legalized in September 2005 and was realized in two hundred most backward districts of the country since February 2006[3].

A. MGNREGA scheme

Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) is a pioneer scheme for providing minimum hundred days of job to the rural and tribal population, whose living merely depends on the daily wages. Family members inhabiting in the same village whose age is above 18 years has to register themselves for the job card, for enlisting themselves in this scheme. Distinctive job card is dispensed to each micro family. Subsequently work will be granted in a period of 15 days, else unemployment benefit is to be provided[4].

MGNREGA is an act with a potential socio-political importance for the rural poor that correspond to the 73rd Amendment. The bill aspires for "at least 100 days of guaranteed employment at the accredited minimal emolument" to adult members of each and every rural household who enlist to do intermittent physical work. For this, a dedicated National Employment Guarantee Fund is to be set up that will be utilized entirely for the enactment of the act. It is inadequate in that the right has been restricted to households, instead of opening it up to every person impoverished, particularly considering intra-household gender discretion.

III. SIGNIFICANCE OF THE ACT

The act aims to improve people's living on sustained basis by developing economic and social infrastructure in rural areas. It directly addresses the reasons of chronic poverty such as drought, deforestation and soil erosion. Rural Employment Guarantee Scheme is demand-driven rather than being supply-driven .

The scheme emphasises on:

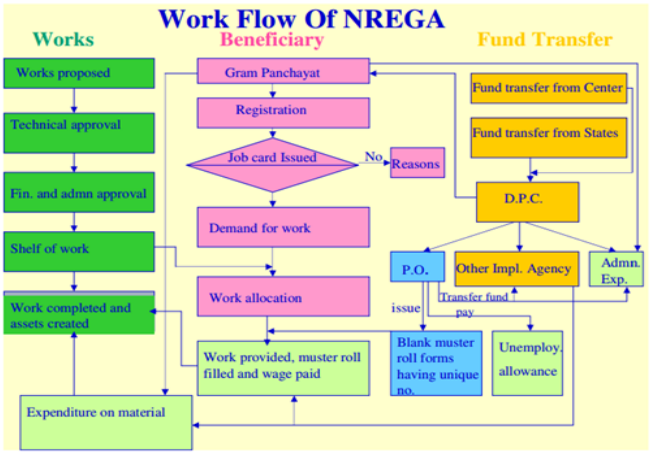
- Water conservation and water harvesting;
- Drought proofing (includes afforestation and tree plantation);
- Irrigation canals inclusive of micro and minor irrigation works; Provide irrigation facility to land owners of households belonging to the SC and ST category or to land beneficiaries of land reforms or that of the recipients under Indira Awas Yojana;
- Refurbishment of traditional water bodies which includes desalting of tanks;
- Land conglomeration;

- Flood control and protection works which also comprise drainage in water logged regions;
- Rural connectedness to make arrangements for all-weather access;
- Some more assignments which may be apprised by the Central Government in dialogue with the State Government[5];

The act has numerous distinguished features. Remunerations have to reimburse every week and in any situation sooner than a fortnight. In occurrence of any detainment in the payment of wages, labourers will be qualified for reimbursement in accordance with the Payment of Wages Act. It is given that under no case "shall there be any differentiation based on gender in the allocation of employment or the remittance of wages, in consonance with the provisions of the Equal Remuneration Act 1976". There are pro visions for compensation and treatment in case of injury and for on-site safe drinking water, care of small children, periods of rest and a first-aid kit. The scheme outlaws the utilization of contractors and labours disburse arrangement. At the minimum of 60 per cent of the expenses of any proposal has to be on reimbursement. These are the provisions that often contravene in innumerable regions of rural India that their magnitude cannot be adequately highlighted.

IV. Workflow of The Act

Adult representative of every single rural household who are keen to do casual manual work at the authoritative nominal wage will have to request to the gram panchayat for enrolment. The gram panchayat will enlist the household, after the required queries have been done, provides the job card comprising particulars of its adult members including their photographs. The filing will be for duration not lesser than 5 years, and may be prolonged every now and then. Recruitment is to be granted to each and every enrolled individual in no more than 15 days of reception of request. Request applications must be for at least 14 days of sustained labour. The gram panchayat is constrained to receive proper appeals and to deliver a dated acknowledgement to the contender. Moreover collaborative applications could be accepted. Candidates who are granted with employment will be granted with written information, with the help of a message dispatched to the locality stated in the job card and through a public intimation presented at the gram panchayat division. As much as attainable, job will be granted inside the limit of five km. Even though the work is granted over the distance of five kilometres , still it is offered in the range of block, and workers will be recompensed with additional 10 per cent of the minimal payment every day , to satisfy supplementary transport and living expenditures.



Source: Official web site of MGNREGA (www.nrega.gov.in)

Fig .2 Workflow of MGNREGA scheme

V. Literature Survey

Several works have been carried out in the past years to explore and analyze the success and failure of this scheme. Different methods have been adapted by different authors, each having its own merits and disadvantages. Some of the previous works carried out are presented in this section.

1) G. Sugapriyan et al [2015] analyses the various factors involved in the successful enactment of the scheme. The study compares the progress of last three financial year data. This scheme is potentially a successful and well designed for bringing the people above poverty line and ensure a financial support for their family to enhance the livelihood and financial state of the rural population. This study shows the success of MGNREGA in Kanchipuram district, employing data mining in addition to the comparison of preceding years statistical data provided by the Government.

2) P. Sumithra et al [2015] analyses the performance of the scheme in Vishakhapatnam by comparing the data of last four financial years. The study shows that the scheme is potentially successful in only few villages of Vishakhapatnam by comparing the progress of last four financial year data.

3) G. Chandra [2015] examines the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA) and its repercussions on the Indian society. The paper also analyse the corruption in MGNREGA scheme as in many public programmes and in countries at all phases of development.

4) Dr.M.Usha Rani [2012] analyses the scheme in Andhra Pradesh .This paper discover the useful pattern on the data of caste wise households which are registered and that are working in NREGS. Rapid Miner data mining tool is used to discover the interesting patterns. Rapid Miner detects clusters based on characterization, Discrimination, and Classification on linear growth rate of households working and wages paid

to them. The results are interpreted in figures. The outcome of this work is useful to the Government to take the decision.
5) M. Ravindar [2016] concentrated on the influence of MGNREGA on women empowerment in the Warangal district of Telangana state. It is deduced that the MGNREGA should be executed in its true intention by redressing blunders in its accomplishment at all levels for attaining intent of the scheme in viable manner.
6) Rahul Bahuguna et al [2016] analyze the influence of MGNREGA on overall economic and social development of beneficiaries in Rudraprayag district of Uttarakhand. The study was carried out in the disaster affected areas of Rudraprayag with beneficiaries as respondents. The result found that the MGNREGA has significantly improved their social and economic well-being.
7) Sumit Garg et al [2013] concentrates on relative study of several data mining techniques and algorithms and to apply appropriate algorithm on educational dataset.

VI. DATA MINING IN E-GOVERNANCE

Government can enhance e-governance strategies more efficiently with the help of data mining techniques. With the incorporation of this strong and powerful technique, government can change its way of conducting various G2G, G2C andG2B services. Government can extract right information from records as per the definite requirement and can use data mining techniques for pattern discovery. Some prominent application areas as per different techniques are discussed below.

A. Classification for customized E-Governance services:

The classification techniques may be employed in E-Governance with different features based on given information for different profiles. A citizen profile mainly includes demographics like Gender, Age, Occupation/job, marital status, etc. This information can help in developing personalized E-Governance services, and also to realize the necessity of consumers, so that government could provide exact services to right consumers.

1) **Decision tree:** Decision tree classifies the data based on series of questions and these questions depend on features that are associated with data. At each and every development of decision tree a parent node contains a question that divide the child node data into possible answers. This hierarchical structure continues until some conclusion is recorded at child node. The decision tree can be employed in any of the e-governance sector like economic sector to check total GDP per capital, for calculation of health issue in different age groups.

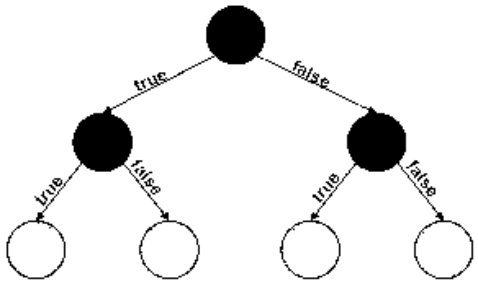


Fig. 3 Decision Tree

Advantages of decision tree:

- Reduced error rate.
- Decomposition is facile.
- Easily operated.

Demerits of decision tree:

- It produce fallible outcome when excessive classes are used.
- A small change in data may alter the decision tree.

2) **Bayesian Classification:** Bayesian classifier is statistical classifier. It is used to project class membership probabilities. This probability about the tuple that belongs to the particular class or not. Bayesian classification is based on Bayes theorem. Assume that X is a data tuple and is observed as "evidence". H is the hypothesis that the data tuple X is associated to a particular class C. We also calculate P(H/X)the probability that the hypothesis H remains valid provided the evidence or observed data tuple X. P(H/X)is the posterior probability of H conditioned on X. Bayes theorem gives a competent approach to calculate the posterior probability, P(H/X) from P(H) , P(X/H)and P(X). Bayes theorem is as follows:

$$P(H/X)=P(X/H)P(H)/P(X)$$

For example, the data tuple is confined to the University described by attribute placement and results and X is a University with placements are good and results are average. Suppose H is a hypothesis that the university gets A grade from UGC. Then P(H/X)is the probability that University X will get A grade and results and placement are known. In contrast P (H) is the prior probability of H. For example, this probability that any given University will get A grade, regardless of placement and results. The posterior probability P(H/X)is based on more information about the University in comparison of prior probability P(H) , which is independent of X. Similarly, P(X/H) is the posterior probability of X conditioned on H. That is University has good placement and average results, and University will get A grade.

Advantages of Bayesian classification:

- Provides preciseness and speed to large datasets.
- Least possible error rate.
- Manages the streaming data flawlessly.

Disadvantages of Bayesian Classification:
It is less accurate since it is feature independent.

B. Clustering on E-Transactions:

Clustering is a technique which divides data into classes having different characteristics and data within every group having similar characteristics. Clustering helps to divide e-governance data and develop some useful patterns on the basis of which government can draw definite decision for data presented in one type of cluster. For example, clustering may be applied on e-transaction rate on different Indian states which will divides data into cluster having set of states with similar e-transaction rate. This could help in preparing similar strategies for high transaction oriented states and low transaction oriented states respectively.

Clustering algorithms can be classified as partitioning methods, hierarchical methods, density-based methods, grid based methods, and model-based methods, k-means algorithm, graph based model etc.

1) **K means algorithm:** K-means algorithm is a procedure of cluster analysis that partitions *n* observations into *k* clusters in which each observation stands to the cluster with the closest mean. K-means algorithm [also denoted as Lloyd's algorithm**]** is an uncomplicated iterative procedure to split a given dataset into user prescribed number of clusters, *k*. The algorithm operates on a set of *d* dimensional vectors, $D = \{\mathbf{xi} \mid i = 1, \ldots, N\}$, where **xi**_*d* denotes the *i*th data point. The algorithm initialize by selecting *k* points in _*d* as the initial *k* cluster representatives or "centroid".The initial seed selection covers sampling at random from the dataset and setting them as the solution to cluster, a compact subset of the data influencing the global mean of the data *k* times. Subsequently the algorithm reiterates between two steps until convergence:

Step 1: Data Assignment- Each and every data point is allocated to the nearest centroid, with ties broken unsystematically which results in division of the data.

Step 2: Relocation of "mean"- Each of the cluster representative is displaced at the centre (mean) of all data points designated to it. If the data points seem to appear with probability measure or weight, then the relocation is to the expectation (weighted mean) of the data partitions. The algorithm approach when the assignments (and hence the **cj** values) no longer modify. Every iteration needs $N * k$ comparisons, that diagnose the time complexity of single iteration. The number of iterations essential for convergence differs and may be conditional on *N*, and in first impression, the algorithm is regarded as linear in the dataset size[6].

Merits of K-means algorithm:

- Comparatively efficient and convenient in implementing.
- Aborts at local optimum.
- Can be employed even on large datasets.

Demerits of K-means algorithm:

- Numbers of clusters have to be identified in advance.
- Inadequate to oversee outliers and noisy data.

C. Association Rule Mining: Association rule mining is invention of association relationships or correlation among a group of items. Association and correlation is used to find the frequent item set among large data sets. Association rules are *if then* statements that look for uncover relationship between unrelated data in the relational database.

Merits of Association rule mining:

- Detects sequential patterns.

Demerits of Association rule mining:

- If the support and confidence are not appropriate, eventually association rule is incompetent.

VII. COMPARATIVE ANALYSIS OF ALGORITHMS

This section presents the comparative analysis of different data mining techniques and algorithms which are discussed above. The comparison of algorithms is done based on their merits and demerits and is reviewed in table1.

Table1: Comparison of Algorithms

Algorithm	Merits	Demerits
Decision tree algorithm	• Reduced error rate. • Provides good result with small size tree. • It is easily understood to humans. • Outcome is not influenced by outliers.	• It produce fallible outcome when excessive classes are used. • Small change in data may alter the tree completely.
Naive Bayes classification	• Provides preciseness and speed to large datasets. • Manages streaming data	• It presumes feature independence.

	flawlessly. • It is proficient to deal with real and discrete values.	
K-Means algorithm	• It is rapid. • It is uncomplicated and robust. • It delivers best consequence when data sets are distinct.	• It is insufficient in dealing with nonlinear data sets. • It cannot manage noisy data and outliers.
Association rule	• Developed to detect sequential patterns.	• If the support and confidence are not appropriate, eventually association rule are incompetent.

VIII. CONCLUSION

The Indian society is afflicted with intense social and economical injustice; the NREGA have become a prime new channel for emancipation of rural and tribal population of India. Yet, success of the act will rely on participation of women, tribal population and the poor. The act ameliorates the socio-economic ambience of rural population. Data Mining comprises of many techniques which includes clustering, classification, association rule mining which are considered in the paper with their merits and demerits.

REFERENCES

[1]. Hardik Gohel, "*Looking Back at the Evolution of the Internet*", CSI Communications - Knowledge Digest for IT Community, Vol.38, Issye.6, pp. 23-26, 2015.

[2]. Nikita Jain, Vishal Shrivastava , "*Data Mining Techniques: A Survey Paper*", International Journal of Research in Engineering and Technology, Vol.2, Issue.11, pp.2319-1163, 2013.

[3]. AKB. Kote, "*role of mahatma gandhi national rural employment guarantee programme inrural –urban migration -a gram panchayat village level study in gulbarga districtof karnataka state*", Journal of Global Economy, Vol.7, Issue.4, pp.275-291, 2011.

[4]. P. Sumithra, VV. Kumari, "*Performance Analysis of MGNREG Scheme using Classification*", International Journal of Science and Research, Vol.4 ,Issue.10, pp.223-226, 2015.

[5]. Dr.M.Usha Rani, "*Analysis on Households Registered/Working through Data Mining Techniques on NREGS (National Rural Employment Guarantee Scheme) Data of Andhra Pradesh*", International Journal of Engineering and Innovative Technology (IJEIT), Vol.2, Issue.2, pp.1-10, 2012.

[6]. T. Smitha, V. Sundaram. , "*Comparative Study Of Data Mining Algorithms For High Dimensional Data Analysis*", International Journal of Advances in Engineering & Technology, Vol.2, Issue.4, pp.44-52, 2012.

[7]. Sumit Garg , AK. Sharma, "*Comparative Analysis of Data Mining Techniques on Educational Dataset*", International Journal of Computer Applications, Vol.74, No.5, pp.1-7, 2013.

[8]. Dr.M.Usha Rani, "*Expenditure Analysis Through Data Mining Techniques on NREGS(National Rural Employment Guarantee Scheme) Data of Andhra Pradesh*", IRACST – Engineering Science and Technology: An International Journal, Vol.2, No. 4, pp.122-129, 2012.

[9]. G. Sugapriyan, S. Prakasam, "*Analyzing the Performance of MGNREGA Scheme using Data Mining Technique*", International Journal of Computer Applications, Vol.109, No.9, pp.123-127, 2015.

[10]. Lavannya Varghese, Christina Joseph, Vince Paul, "*Survey on Mining Educational Data and Recommending Best Engineering College*", International Journal of Science, Engineering and Computer Technology, Vol.6, Issue.1, 66-69, 2016.

[11]. Moksha Shridhar, Mahesh Parmar, "*Survey on Association Rule Mining and Its Application*", International Journal of Computer Science and Engineering, Vol.5, Issue.3,pp.129-135, 2017.

[12]. Mala Bharti, Vineet Richhariya, Mahesh Parmar, "*An Implementation of IDS in a Hybrid Approach and KDD CUP Dataset*", International Journal of Research Granthaalayah (IJRG) Indore, Dec-14, Vol. 2, Issue.2, pp. 2-12, 2014

[13]. Mala Bharti, Vineet Richhariya, Mahesh Parmar, "*A Survey on Data Mining based Intrusion Detection Systems*", International Journal of Computer Networks and Communications Security, Vol. 2, No.12, pp. 485-490, 2014.

[14]. NA. Shiekh, MA. Mir, "*Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA): A Right Based Initiative towards Poverty Alleviation through Employment Generation*", International Journal of Science and Research (IJSR), Vol.5 Issue.1, pp.1344-1348, 2016.

[15]. P. Chakraborty, "*Evaluation of National Rural Employment Guarantee Act in Tamil Nadu*", Indian Institute of Technology, Madras, pp.1-29, 2010.

[16]. M. Dey, S.S. Rautaray, "*Disease Predication of Cardio- Vascular Diseases, Diabetes and Malignancy in Lungs Based on Data Mining Classification Techniques*", International Journal of Computer Sciences and Engineering, Vol.2, Issue.4, pp.82-98, 2014.

[17]. Bhavesh Patankar, Vijay Chavda, "*A Comparative Study of Decision Tree, Naïve Bayesian and k-nn Classifiers in Data Mining"*, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4, Issue.12, pp.24-31, 2014.

