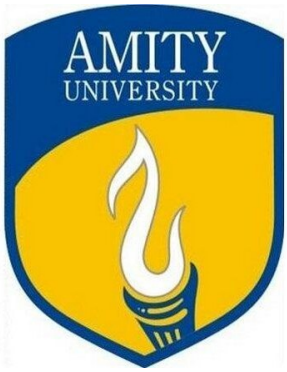

AMITY UNIVERSITY,RAJASTHAN

THE MANAGEMENT OF MISSING DATA IN DATA MINING

SUBMITTED BY:RIYA TYAGI

2nd year student,M.Sc(Data Science)

ABSTRACT:

Data is usually incorrect. The topic of categorization with insufficient data is novel. This study provides a categorization for partial survey data. The process of classification using incomplete data is complex, and its effectiveness is decided by the technique employed to deal with missing data. Missing data occurs when no data value for an attribute or feature in a dataset is retained. This study provides a high-level overview of incomplete data handling methodologies and analyses the various classification and missing data strategies. It presents a number of classification techniques that take advantage of incomplete data.

KEYWORDS: missing data, data handling, techniques

INTRODUCTION:

Data mining, or extracting hidden predictive information from massive databases, is a promising tool for forecasting future trends and behaviors that can provide significant benefits and which are incapable of predicting future behavior and cannot be recreated on new data sets, a little amount of usage It helps businesses to make data-driven, proactive choices. It's an interdisciplinary project. Database technology, artificial intelligence, machine learning, and neural networks have all contributed to this field. Knowledge-based systems, knowledge acquisition, information, networks, statistics, pattern recognition Data visualization, high-performance computing, and retrieval incomplete data can have a big influence on the outcome. The performance of algorithms and the quality of learned patterns As a result, how do you deal with missing data? Data is a vital and challenging subject to handle in the fields of machine learning and data mining.

There are numerous approaches for dealing with incomplete data in terms of categorization.

Incomplete Data Types: Little and Rubin identify a set of missing mechanisms that are largely acknowledged by the scientific community. Missing data can arise due to one of three mechanisms

1. **Missing completely at random (MCAR):** MCAR is the chance that an observation (Xi) is missing, is independent to the value of Xi or any other variable, and the reason for missing is purely random. MCAR occurs when a tube carrying a blood sample of a study participant is inadvertently damaged (resulting in the blood parameters being unable to be assessed) or when a questionnaire of a research subject is mistakenly misplaced. This is an uncommon occurrence in the actual world and is generally explored in statistical theory.
2. **Missing at random (MAR):** MAR is the likelihood of the observed pattern given the observed and unseen data, and it is independent of the values of the unobserved data. A participant omitting an answer on a questionnaire is an example of this. This approach is widely used in practise and is typically regarded as the default form of missing data.
3. **Not missing at random (NMAR):** If the likelihood that an observation is missing is determined by information that is not observed, the missing data is referred to as NMAR. High-income individuals, for example, may be more hesitant to give income information. This is a tough problem, and there is no general solution.

REVIEW OF TECHNIQUES FOR LOST DATA:

There are four commonly used methods for dealing with missing values:

1. Removing the tuples: If any of the data's factors has a missing value, the tuples are eliminated.Take the full observation out of the equation.

2. Manually filling in the missing value: In a real-world situation, a huge data collection with many missing values, this strategy is time-consuming and unworkable. For example, use a global constant to replace any missing attribute values. a huge data collection with a substantial number of missing values.
3. Use the attribute mean value to fill in the missing estimations of a variable. Substitution The metrics of central tendency are mean, median, midrange, (Max + Min)/2, and mode. Missing data can be handled in a variety of ways; a few examples are shown below.

TECHNIQUES IGNORED BY MISSING DATA

1).LISTWISE DELETION : If a case has missing data for any of the variables, remove it from the analysis using listwise deletion (or complete case analysis). It is the default setting in the majority of statistical applications.

2).PAIRWISE DELETION:is a feasible case technique (PD). This technique analyses each attribute independently. For each feature, all recorded values in each observation are evaluated, and missing data is disregarded.

MISSING DATA IMPUTATION TECHNIQUES

The imputation method is a type of methodology that attempts to fill in missing information with approximated values. The idea is to use known relationships that can be identified in the data set's valid values to aid in guessing missing values. This field is concerned with the imputation of missing data.

MEAN VALUE IMPUTATION METHOD: One of the most commonly utilised approaches is mean imputation. It consists of replacing missing data for a specific component or attribute with the mean of all known values of that attribute in the class to which the missing attribute instance belongs.

HOT DECK IMPUTATION (HD): Given an incomplete pattern, HD fills in the missing data with values from the input data vector that are closest in terms of the qualities that are known in both patterns. By replacing distinct observed values for each missing, HD aims to safeguard the distribution. Cold deck imputation is a comparable HD approach that uses a different data source than the present dataset.

K-Closest NEIGHBOR IMPUTATION (KNN): This technique estimates and replaces missing data using k- nearest neighbour algorithms. The key advantages of this method are as follows: a) it can estimate both qualitative and quantitative qualities; and b) it is not necessary to build a prediction model for each attribute with missing data.

K-MEANS CLUSTERING METHOD: K-Means classifies or groups objects based on attributes/features into k groups. The grouping is completed by reducing the sum of squares of distances between the data and the cluster centroid. It gives a rapid and accurate method of calculating missing numbers.

FKMI (FUZZY K-MEANS CLUSTERING IMPUTATION): The membership function is very significant at FKMI. Every data object has a membership function that indicates how much the data object belongs to a specific cluster. Data objects would not be assigned to a specific cluster, as indicated by the cluster's centroid (as in the case of K means), due to the varying membership degrees of each data with all K clusters.

REGRESSION IMPUTATION: The values from the features are observed using the regression technique for imputation, and then predicted values are utilised to fill in the missing values.

MULTIPLE IMPUTATIONS: Because the imputed values are drawn from a distribution, they are naturally variable. Thus, multiple imputations (MI) highlight the limits of single imputation by introducing a new type of error based on variance in parameter values throughout the imputation, known as between imputation error. It substitutes two or more acceptable values for each missing component, providing a distribution of alternatives.

MODEL-BASED TECHNIQUES FOR DATA LOSS

Maximum likelihood strategies are used to infer the parameters of a model constructed for all data. Maximum likelihood techniques that employ variants of the Expectation Maximization algorithm may handle parameter estimation in the case of missing data.

MAXIMUM PROBABILITY: We may use this method to create the variance-covariance matrix for the variables in the model based on all available data points, and then utilise the variance-covariance matrix to estimate our regression model.

EXPECTATION-MAXIMIZATION ALGORITHM (EM): It is based on a stage of constant expectation and maximisation that is repeated many times until maximum likelihood estimates are obtained. It demands a large sample size and random missing data (MAR).

V. CONCLUSION:

The primary goal of this work was to discuss the various approaches to categorization when dealing with incomplete data values.Many real-world applications of pattern categorization suffer from missing or insufficient data. Data may contain unknown features due to a variety of factors, such as sensor failures that result in a warped or distorted image. Non -response in surveys, unquantifiable value, data blockage by noise. Missing data handling has become a need for pattern classification because an incorrect categorization can lead to serious consequences. Missing data management might lead to significant errors or inaccurate classification findings. It can be demonstrated that statistical and numerical data are both relevant. Methods and machine learning approaches have shown some progress in the issue arena. This is currently being contested. Machine learning-based artificial neural networks are being evaluated as missing data imputation solutions. The K-Nearest Neighbor technique and Self-Organizing Maps are two popular algorithms (SOM). SOM variants such as tree-structured SOM are also available. The EM algorithm is widely utilised.

REFERENCE:

1).Swati Jain,Mrs Kalpna Jain and Dr Naveen Chodhary "A Survey paper on missing data in data mining"INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT](2016)

2)Jehanzeb R. Cheema "A Review of Missing Data Handling Methods in Education Research" Review ofEducational Research(2014)

3)Dr.C.Yamini and M.Kowsalya "SURVEY ON CLASSIFICATION OF INCOMPLETE DATA HANDLING TECHNIQUES" "INTERNATIONAL JOURNAL OF ADVANCE RESEARCH ANDENGINEERING(2015)

4)Akshaya R, Anushree G ,Devaki Sai Mahitha ,Madhura and Shylaja B "HANDLING MISSING DATA : MICE (A DATA MINING APPLICATION)"Department of Computer Science,DSATM , Karanataka(2021)

5)Devi Priya R and sivaraj R, "A REVIEW OF MISSING DATA HANDLING METHODS" InternationalJournal On Engineering Technology and Sciences – IJETS(2015*)*