Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Prediction Of Misclassification Data Based On Cognitive Computation Approach (Cca)

S. Kanchana

Dept. of Computer Science, FSH, SRM Institute of Science & Technology, Kattankulathur 603203, Chennai, India

ABSTRACT

In quantitative analysis, missing data create unavoidable problem in real world large datasets. Due to the issues the conclusion of the computational process cause bias outcome, increasing rate of error data, and more inconvenient to attain the process of imputation. Prediction model is one of the elegant methods for managing missing data. This article introduced, the most powerful approaches for the prediction of misclassification data using Machine Learning (ML) techniques. Also it explores the study of Adaptive Computation and Pattern Knowledge Theory using effective Cognitive Computation Approach (CCA). Several strategies describe the classification of predictive techniques using efficient Supervised Machine Learning Algorithm. Main goal is to provide general guidelines on selection of suitable data imputation algorithms and also implementing Cognitive Approach in Machine Learning Techniques. The proposed approach generated more precise, accurate results than the other predictive approaches. The Experimental results performed both real and synthetic dataset, proved that the proposed approach offers valuable and optimistic insight to the prediction of misclassification information.

KEYWORDS: Cognitive Computation Approach (CCA), Machine Learning Techniques (ML), Misclassification Data, Predictive Techniques, Supervised Machine Learning Algorithm.

***Corresponding author**

Dr. S.Kanchana

Department of Computer Science

Faculty of Science and Humanities

SRM Institute of Science and Technology

Kattankulathur – 603203,

Email: kskanch@gmail.com Mob No- 9444468583

INTRODUCTION

Data analysis has various aspects and methods, which includes apparent approaches with a different domain. Data analysis particularly specifies data mining techniques which focus on modelling and knowledge discovery. The raw facts of data include operational or transferable data, non-operational data, and Metadata that can be handled by a computer which provides information. Information can be converted, exchanged, cooperated and designed into knowledge as per historical design and future trends. The preliminary pattern for data analysis is nothing but Data integration, which is closely related to data modelling. Data mining or knowledge discovery is scrutinizing raw facts of data from distinct aspect and rehashing it into valuable knowledge. Data mining[1] is a powerful automation tool with high capabilities to assist the operation which emphasizes the salient facts of data accumulating around the attitude of the client. It determines the information within the data so that inquiries and description cannot efficiently expose it. Knowledge discovery in the database is the process of computer-aided mechanically of drilling over the data, determining huge firm of data and then deriving the context of data from the databases. Data mining process is deployed by the organization to produce powerful and functional data from unprocessed data. By using mining techniques, business people acquiring more knowledge about buyers and expand more effectual promoting approach, escalating sales and declining costs. Data mining techniques performed based on efficient information collection, warehousing and computer processing. Supermarket, beauty parlour, textiles showrooms are widely known consumer of data mining techniques. Many stores offer free membership cards to customers that provide them to approach reduced amount not applicable to non-members. Membership cards enables easy for stores to trail the details of shopping, who is purchasing, when they purchase, and at what price. After analysing it, the stores can use the data for various objectives such as offering customers vouchers targeted to purchasing habits and when to insert and close item on sale at full price.

Data mining is generally used contemporarily by companies with powerful services, with focus on trade, economy, transmission, and commercial management. It authorizes these companies to regulate communication among internal aspects like cost, product positioning, or staff skills and external aspects like financial indicators, contest, and customer enumeration. Mining technology allows them to fix the impact on marketing, client satisfaction, and corporate benefits. Finally, it implements them to focus on summary information to view structured transactional data.

LITERATURE REVIEW

Little and Rubin[2] summarize the mechanism of imputation method. Also introduces mean imputation[3] method to find out missing values. The drawbacks of mean imputation are sample size is overestimated, variance is underestimated, correlation is negatively biased. For median and standard deviation also replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. Different types of machine learning techniques are supervised and unsupervised machine learning techniques[4] summarized in. Classification of multiple imputation and experimental analysis[5] are described in Min Pan et al.[6] summarize the new concept of machine learning techniques like NBI also analysis the experimental results which impute missing values. Comparisons of different unsupervised machine learning techniques are referred from survey paper[7]. To overcome the unsupervised problem Peng Liu, Lei Lei et al.[8] applied the supervised machine learning techniques called Naïve Bayesian Classifier.

MISSING DATA ANALYSIS

Data imputers replace the missing values with the help of fake values. General statistical methods implemented by the programmers remove raw facts of data and then leads to process with the rest of the data. However, this experimental analysis produced biased results of the parameter. The common statistical report fails to produce the constant result for the missing data issues. As an alternate process, the imputers proceed with multiple imputations which were proposed by Rubin (1987) to encounter this issue. To overcome this issue, imputation technique takes a major role because the analyst may not know the necessary information to overcome the issue that appears in the presence of missing values in the data sets, since it is the responsibility of imputers to handle the missing data problems in terms of multiple imputation techniques[9,10]. The following fig. 1 depicts the process of missing data analysis. Large data set consist of η number of missing values and to overcome such issues can ignore the missing data. Ignoring the missing data in the large data sets causes bias result. The imputation process takes the major part to avoid the bias effects, studying the various analysis processes statistical approach to allow for leaving out data and machine learning approach based on predicted values within each dataset. Analysis results of each data set to pool for final estimation.

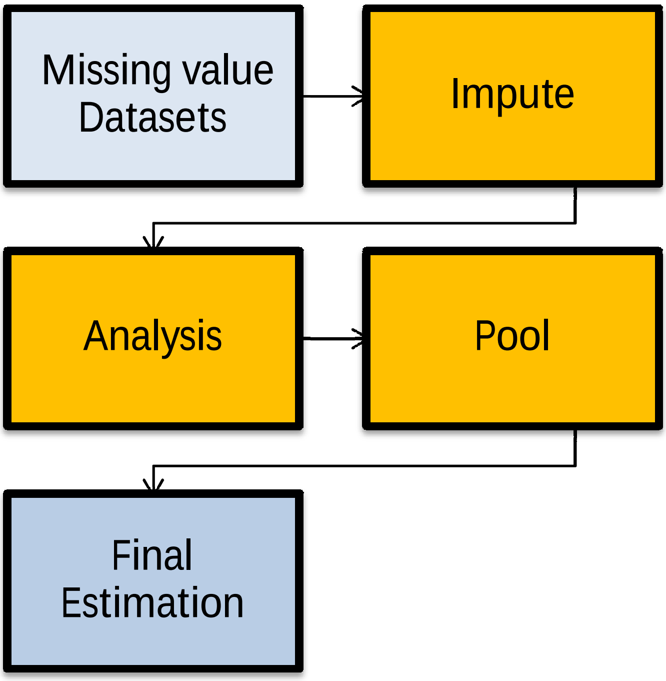


Figure 1. Process Of Missing Data Analysis

It is a statistical approach for dealing with incomplete data which estimates the valid result. An extendable version of single imputation is by replacing every incomplete data by a η group of imputed values, which are extracted by reasonable circulation. Analysis part handles the η sets of imputations to fill the incomplete data η times, which is allocated in the pool for further estimation to produce η complete data sets in the absence of incomplete information. Later statistical analysts combine the analysis results in data with each and every η completed data set which produces the assured final estimated inference by using multiple imputation rules proposed by Rubbin (1987).

To improve the performance of accuracy rate, apply predictive data analysis technique to classify the data. The proposed work develops the complete assurance and it offers an analytical thinking of existing approaches for the major technical and theoretical issues were undertaken. The sequence of recent techniques[11] can improve the behaviour of traditional classifiers in terms of an incomplete sequence of event. Incomplete data has proved to be the rule rather than an exception in real-world data mining problems. At the same time, it represents a challenge for achieving a successful data mining process, since statistical techniques have not been designed to deal with incomplete data and most of them employ straightforward and inefficient approaches. Consider all missing values as a unique value; replace all missing values with NULL, remove all instances/attributes having missing values.

CLASSIFICATION OF PREDICTION TECHNIQUES

The major techniques for handling the incomplete data consider the easiest term for the analysis of missing values imputation problem[12], which simplify the data set in terms of removal of all models with specific values. At last, imputation of missing value problem can be rectified by various imputation techniques. Unexpectedly, the imputation techniques are approachable only for missing values motivated by missing completely at random and few of them are eligible for missing at random mechanism. In some cases, if the missing values are not raised by not missing at random mechanism, then it must be dealt with the cause of data, and the relative pattern of missing data mechanism must be taken into an explanation.

Generally machine learning techniques are analysed in different categories in terms of predicting the incomplete values of large datasets, which propose outstanding performance or response from learning system. Basically, the three objective learning techniques fig.2 in machine learning classifications are supervised, unsupervised and reinforcement learning techniques, which propose to perform the accurate predictions depending on the prior observations based on the classification problem[13,14]. To handle with incomplete data imputation issues, there is no specific method to implement the imputation process in data analysis. The main advantage of these techniques is to generate more accurate data analysis without human expertise.

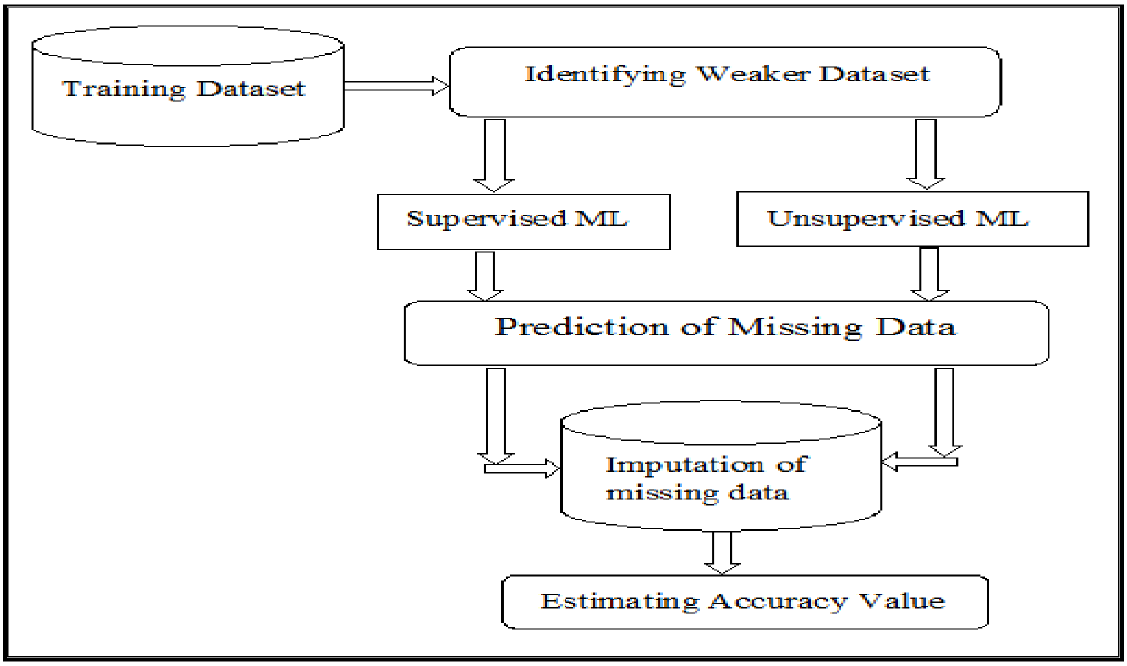

Training Dataset
Identifying Weaker Dataset
Supervised ML
Unsupervised ML
Prediction of Missing Data
Imputation of missing data
Estimating Accuracy Value

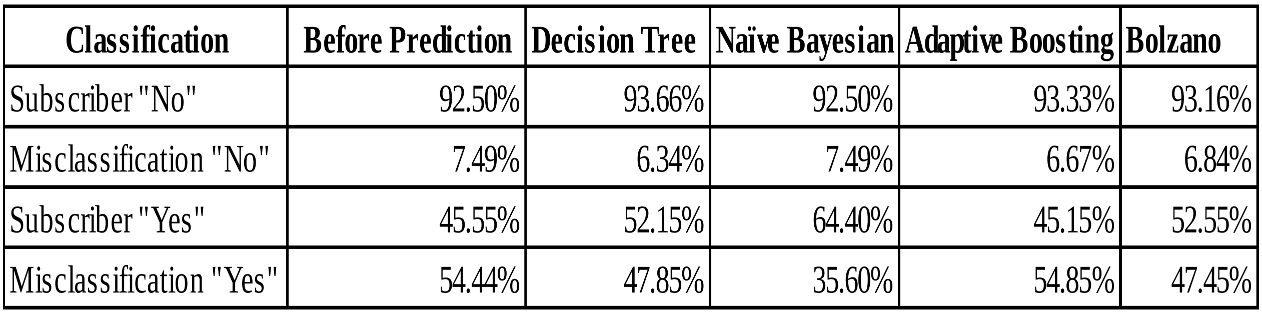
Figure 2. Existing System

RESULT ANALYSIS

The proposed research work introduces new, efficient approach, namely CCA – Cognitive Computation Approach which generates easy execution process to predict the missing values using machine learning techniques and pattern knowledge theory, which is very simple to understand, and produces high rate of accuracy compared to other techniques. Cognitive approach provides us with more idea for further research in order to understand better the present phenomenon. To reduce the number of false positives and the number of false negative values and to find the limit of missing values, the proposed approach requires a different mathematical approach to implement the lower and upper bound algorithm using Bounded Monotone Sequence theorem. To find the interval of transactions between the limit of unknown data, the Bolzano Weierstrass theorem will satisfy the requirement of CCA techniques to validate the prediction in large samples of the dataset. Cognitive approach plays an important role in perspectives process because it can easily recognize other perspective knowledge, which acquires moderate pattern recognition capacity.

Considering prediction model analysis, we used Decision Tree (DT), Naïve Bayesian (NB), Adaptive Boosting (ADAB) and mathematical approach of Bolzano Weierstrass Theorem (BWT) compared with a dataset of Bank Marketing for the validation of accuracy in percentage. The below Table 1. states the percentage value of accuracy using various models.

Table 1. Percentage Value Of Accuracy Using Various Models



Classification	Before Prediction	Decision Tree	Naïve Bayesian	Adaptive Boosting	Bolzano
Subscriber "No"	92.50%	93.66%	92.50%	93.33%	93.16%
Misclassification "No"	7.49%	6.34%	7.49%	6.67%	6.84%
Subscriber "Yes"	45.55%	52.15%	64.40%	45.15%	52.55%
Misclassification "Yes"	54.44%	47.85%	35.60%	54.85%	47.45%

Through the analyses descirbed above, we have been able to show that the work carried out by the Cognitive Computation Approach (CCA) model has greater predictive capability and accuracy than the popular prevelant models. In both the “Miclassification NO” and “Subscriber YES” cases, the CCA models emerged superior. In addtion, it is important to note an additional formidable capability of the CCA model, which is that it can also adapt using the Cognitive Pattern Knowledge Theory to predict the misclassification rate of hidden attributes in large dataset.

CONCLUSION

The existing research work was performed in order to predict the misclassification of data in a large dataset using effective supervised machine learning approach. The limitations of the existing research do not carry out the assessment of cognitive approaches to all the parameters in the dataset to predict the missing values using a supervised model. After analyising the demerits of the existing system,taking out the proposed model, Cognitive Computation Approach (CCA) approach was chosen. Generation of adaptive computation and pattern knowledge theory for the prediction of misclassification data.

REFERENCES

1. Doh-Soon Kwak, and Kwang-Jae Kim "A Data Mining Approach Considering Missing Values for the Optimization of Semiconductor-Manufacturing Processes" *Expert Systems with Applications* 2012; 39: 2590-2596
2. R.J. Little and D. B. Rubin. Statistical Analysis with missing Data, John Wiley and Sons, New York, 1997.
3. R.S. Somasundaram, R. Nedunchezhian, "Evaluation on Three simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values", International Journal of Computer Applications, May 2011; 21(10):14-19.
4. Jeffrey C.Wayman, "Multiple Imputation for Missing Data: What is it and How Can I Use It?" Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL, 2003; 200: 2-163.
5. Mrs.R. Malarvizhi, Dr. Antony Selvadoss Thanamani, "K-Nearest Neighbor in Missing Data Imputation", International Journal of Engineering Research and Development, November-2012,
6. Alireza Farhangfar, Lukasz Kurgan and Witold Pedrycz, "Experimental Analysis of Methods for Imputation of Missing Values in Databases.
7. K. Lakshminarayan, S.A. Harp, and T. Samad, "Imputation of Missing Data in Industrial Databases", Applied Intelligence, 1999.; 11: 259-275
8. Peng Liu, Lei Lei, "Missing Data Treatment Methods and NBI Model", Sixth International Conference on Intelligent Systems Design and Applications, 0-7695-2528-8/06.
9. Yeats Ye, "Multiple Imputation for Survey Data Analysis" *Paper CC-016*
10. Yang C. Yuan, "Multiple Imputation for Missing Data: Concepts and New Development(Version9.0)" https://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf

11. Johannes Schreyer, Christian Geiß, and Tobia Lakes, “TanDEM-X for Large-Area Modeling of Urban Vegetation Height: Evidence from Berlin, Germany” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, May 2016; 9(5)
12. Federico M. Sukno, John L. Waddington, and Paul F. Whelan, “3-D Facial Landmark Localization with Asymmetry Patterns and Shape Regression from Incomplete Local Features” *IEEE Transactions on Cybernetics,* September 2015; 45(9)
13. Samuel H. Hawkins, John N. Korecki, YoganandBalagurunathan, YuhuaGu, Virendra Kumar, SatrajitBasu, Lawrence O. Hall, Dmitry B. Goldgof, Robert A. Gatenby, and Robert J. Gillies, “Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features” *IEEE*, 2014; 2
14. Samuel H. Hawkins, John N. Korecki, YoganandBalagurunathan, YuhuaGu, Virendra Kumar, SatrajitBasu, Lawrence O. Hall, Dmitry B. Goldgof, Robert A. Gatenby, and Robert J. Gillies, “Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features” *IEEE*, 2014; 2