Prediction of Misclassification Data using Cognitive Bayes Computation Techniques (COBACO)

S. Kanchana

Abstract*: Missing data arise major issues in the large database regarding quantitative analysis. Due to this issues, the inference of the computational process produce bias results, more damage of data, the error rate can increase, and more difficult to accomplish the process of imputation. Prediction of disguised missing data occurs in the large data sets are another major problems in real time operation. Machine learning (ML) techniques to connect with the classification of measurement to enforce the accuracy rate of predictive values. These techniques overcome the various challenges to the problem of losing data. Recent work based on the prediction of misclassification using supervised ML approach; to predict an output for an unseen input with limited parameters in a data set. When increase the size of parameter, then it generates the outcome of less accuracy rate. This article presented a new approach COBACO, an effective supervised machine learning technique. Several strategies describe the classification of predictive techniques for missing data analysis in efficient supervised machine learning techniques. The proposed predictive techniques COBACO generated more precise, accurate results than the other predictive approaches. The Experimental results obtained using both real and synthetic data set show that the proposed approach offers a valuable and promising insight to the problem of prediction of missing information.

Keywords* : COBACO, Machine Learning, Prediction techniques, Supervised Machine Learning.

I. INTRODUCTION

Data mining is generally used contemporary by companies with powerful services focus on trade, economy, transmission, and commercial management. It accredited these communities to classify exchanging information between internal aspects like cost, product positioning, or staff skills and external form like financial indicators, contest, and customer enumeration. Data mining techniques admit them to conclude the impact on marketing, client satisfaction, and corporate benefits. Certainly, it implements the authority to focussing into supply information to view structured transferable data. Even though data mining is approximately new description of concept for continuous innovation, the technology is not involved. Data mining [1] is a powerful automation tool with high capabilities to assist the operation which emphasizes the salient facts of data accumulates around the attitude of the client. Data mining process deploys by the organization to produce unprocessed data into powerful and functional data.

This overview provides the most common data mining algorithms which consist of two sections. Each section defines a number of data mining algorithms at a high level of techniques like classical and next generation techniques. Classical techniques [2] consist of statistics, neighbourhoods, and clustering, and next generation techniques reside trees, networks, and rules. These two sections have been split up based on when the data mining technique was established and when it became technically sophisticated enough to be used for trade, especially for helping the development of customer relationship management systems. Also, it helps to understand the uneven differences in the techniques and well equipped enough not to be confused by the vendors of different data mining tools. Statistical techniques are not data mining, which can be consumed by the data and are used to design patterns and build predictive models. Clustering is the process of classification of physical or abstract objects into classes of similar objects and dissimilar objects into another cluster. The advantages of clustering techniques [3] are flexible to make changes and assist single out useful features so that it identify different groups. Learning and classification steps of a decision tree are simple and fast, easy to comprehend and do not demand any domain knowledge.

The Neural network [4] is an interconnected group of nodes, which exchange messages between each other. The most significant advantages can be extremely accurate predictive models that can be utilized across a large quantity of dissimilar types of problems. Neural networks perform learn in a very sensitive but beneath the hood the algorithms and techniques that are being deployed are not absolutely different from the techniques found in statistics or other data mining algorithms.

A. Machine Learning Techniques

Research community proposed various suitable algorithms during the prediction of incomplete data analysis in terms of data mining approach. Machine learning techniques mainly concentrate on the pattern recognition and computational learning theory [5,6] in artificial intelligence which analyses the term and structure of algorithms in order to handle the predictions of missing data. This specific mechanism is based on to frame a structure of algorithms can able to learn and to build predictions of missing data, which is nearly connected to perform the arithmetic operation through statistics analysis. Learning techniques used the various applications in terms of predicting data analysis [7] by

Revised Manuscript Received on January 5, 2020

Dr.S.Kanchana*, Department of Computer Science, SRM Institute of Science and Technology, Kattankular, Chennai, India, Email: kskanch@gmail.com

focussing various basic predictive data analysis techniques in the name of unsupervised learning approach. In parliamentary law to do predictive analysis by means of machine learning and statistical techniques are linked to each other in terms of methodological structure. These entire analysis models provide more innovative information through factual communication to the researchers, developers and system and data analysts regarding recent movement in data. Various classification of learning assignment is available in machine learning techniques includes supervised, unsupervised and reinforcement learning.

B. Classification of Prediction Techniques

Generally machine learning techniques are analysis into different categorize in terms of predicting the incomplete values of large datasets which propose outstanding performance or response from learning system. Basically three objective learning techniques in machine learning classifications are supervised, unsupervised and reinforcement learning techniques which propose to perform the accurate predictions depends upon the prior observations based on the classification problem [8, 9]. The main advantage of this technique is to generate more accurate data analysis without any human expert. Supervised techniques which come under the process of machine learning task for providing specific inference from the identified data set. These techniques estimate various training dataset for producing optimal output result which is required for arranging unknown models. With the help of optimal algorithm will decide for correcting the class specification for identical attributes, which need to optimised label attributes in the desired data for producing the unexpected result. The clustering approach can improve the scalability of machine learning techniques in terms of machine learning algorithm by grouping the related data items in a known training dataset. Multistage clustering techniques which incorporate the imputation of missing values also proposed. Clustering and along with networking algorithms attempt to duplicate incomplete measurements in a dataset by sufficiently manipulating the duplicates using various clustering techniques.

II. LITERATURE SURVEY

Alireza Farhangfar et al [10] focused important issue faced by researcher and practitioners for incompleteness of data, in terms of missing or erroneous values. Suggested different strategies such as deletion of incomplete, and imputation of missing values through various statistical and machine learning (ML) techniques. Edgar Acuna and Caroline Rodriguez [11] proposed to treat missing data and the one used more frequently is deleting instances containing at least one missing value of a feature. Peng Liu, and Lei Lei [12] proposed Naïve Bayesian imputation method (NBI) to handle effective popular missing data treatment model case deletion, parameter estimation, Mean/Mode imputation (MMI), regression, hot deck and cold deck imputation, multiple imputation, K-NN and C4.5 imputation method along with three strategies are Order irrelevant strategy (NBI-OI), Order relevant strategy (NBI-OR), and hybrid strategy (NBI-Hm). Kai Zhang et al [13] presented semi supervised learning to improve the learner's performance by using unlabeled data. Miroslaw Pawlak [14] computed a number of nonparametric kernel classification rules, consistency and speed of convergence of kernel classification rules established from missing data. Prediction density approach, the deletion techniques, and stochastic mechanism of mechanism of generation of missing values can be introduced by imposing probability distribution on the variables.

III. COBACO APPROACH

The proposed approach is COBACO stands for Cognitive BAyes COmputation Approach, which derived from traditional approach to improve the data quality of prediction. In order to determine the sub classification portion of each attributes can estimate by using effective supervised technique called COBACO. The main goal is to predict the value of missing data and impute the data with high quality prediction for training data set. From the training data set, along with all imputed values, generating the frequent table for variable selection. So that can avoid the over-fitting problem occurred in the training data set. Applying cognitive approach for every instances of the data set to prepare high quality of positive and misclassification rate of prediction on the customer. It's a combination of posterior prediction and ensemble approach to improve the quality of the prediction techniques. To improve the framework for learning classifiers with a small value of iteration by increasing the speed limit with all the instances of attribute. Generate sub classification rule for all the instances in the data set, so as to increase the rate of high positive prediction value. This iteration continues for the entire instance of the data set without skipping any attribute from the frequent table. Apply the traditional approach to calculate the accuracy rate in terms of positive prediction, negative prediction, and misclassification rate of positive and negative value. Along with the test data find the true positive accuracy rate of prediction to the training data.

Analyzed few drawbacks from the above prediction techniques in term of accuracy, data quality, error rate and elapse time for the execution. COBACO has strong quality of independence techniques. Can overcome the problem of zero insertion in the empty field so that the prediction value will be more. This method even fit to generate the calculation of precision value and also make more reliable for the estimation of the probability of each instances. Reduce the noisy data and the error rate of both positive and negative misclassification rate. Another advantage of proposed system is no restriction of applying statistical tools. The overview of COBACO framework process is to describe the simple and easiest techniques in all other machine learning models. The process starts with original data in terms of all the attributes in the data set. Generate frequent table which consist of frequently used data items to compute the classification of large data set and to find the positive classification of prediction, negative classification of prediction, positive misclassification of prediction and negative misclassification of prediction from the training data set. Based upon the reduction of data items in a record, can reduce

the unwanted attributes which is not require for the estimation process. The main advantage of this record is to reduce the elapse time for the calculation of positive and negative classification and also utilize the less time for the estimation of error rate. Compare to other prediction techniques COBACO produce high accuracy rate and the elapsed time taken for the process is 0.2 sec higher than Naïve Bayesian techniques.

The cognitive approach is a new process to describe the human activities which concentrate on how the human can think, and explained the processes in which the human can react. Such activities expressed as a set of scientific processes and its controlled by own thought processes. Cognitive approach is a special methodology to rectify the issues when less number of necessary content is available. To implement this approach in terms of experimental analysis by using special programming tool called Bayesian. It is represented by:

E – Explanation of variables and the description in the data set
Q – Query about the distributions
π – Definition of variables, decomposition of the functions and the models.
δ – Description of the data set
V- Variables available in the data set.
D – Decomposition describe the subset variables
M – Models associate the parametric form of distribution

A. Metrical Analysis of COBACO Approach

COBACO approach performs simple interpretation process to predict the missing values of large data set. Compare to existing approach, the COBACO model easily generate more than 10,000 instances of classes along with maximum number of attributes. It support with maximum instances of large training data set. Unlike NB, it's not sensitive to unrelated data and also assumes independence of features. COBACO is more expressive to hold the maximum number of parameters and learning rate. It require very less effort for data preparation and assessment for feature selection. The best feature of this techniques provide easy way of understanding, easy to code and very simple to generate because proposed work generating pattern knowledge theory which solve the complex process. Finally, compare to other technique COBACO perform the sub classification rule using cognitive approach for each and every attribute in the large data set. The following Figure1. Describe the performance of various approach rate of accuracy for positive and negative prediction of unknown data, in which COBACO generates high prediction of positive classification rate and low assumption of negative misclassification rate.

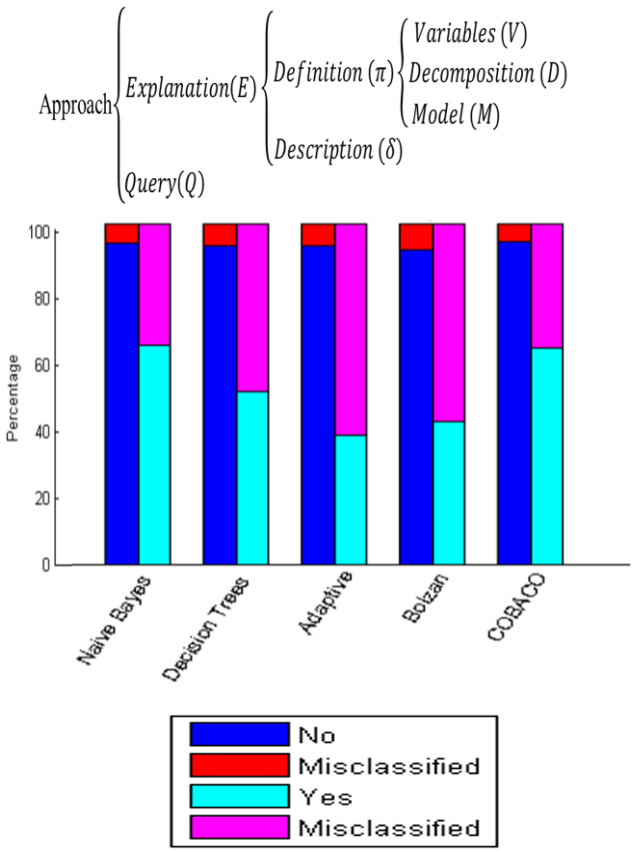

Figure 1. Comparison chart with proposed techniques

The proposed research work introduces new, efficient approach, namely COBACO – Cognitive Bayes computation technique generates easy execution process to predict the missing values using machine learning techniques and pattern knowledge, theory which is very simple to understand, and produce high rate of accuracy compare to other techniques. COBACO approach provides us more idea for further research in order to understand better the presence phenomenon. To reduce the number of false positive and the number of false negative values, 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 COBACO techniques to validate the prediction in large samples of the dataset. Cognitive approach play an important role in perspectives process because it can easily recognized other perspectives knowledge which acquire moderate pattern recognition capacity. One of the incredible problems facing in a large data set is incomplete values. The outcome of the estimation generates more bias results. Implement many algorithms to predict the misclassification data in order to perform the accurate results. Apart from machine learning tools and lots of mathematical theorems, the fundamental equality of mind is the most important instrument for the prediction of misclassification data in terms of the cognitive approach.

IV. EXPERIMENTAL ANALYSIS

Prediction is a process of assuming about unknown value in the datasets using statistical or machine learning approach. Machine learning goal is to predict whether the client will subscribe a term deposit or not. This analysis represented the effective computation techniques in order to predict the missing data. The real data provided by UCI Machine Learning Repository, which was collected from Portuguese Bank, with a total of 45,211 instances and 21 number of real attribute characteristics

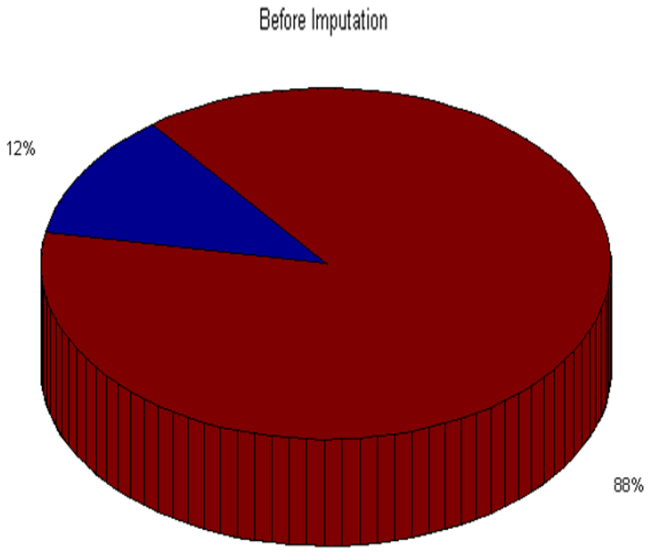

Figure 2. Classified and Misclassified Structure of Dataset Before Imputation

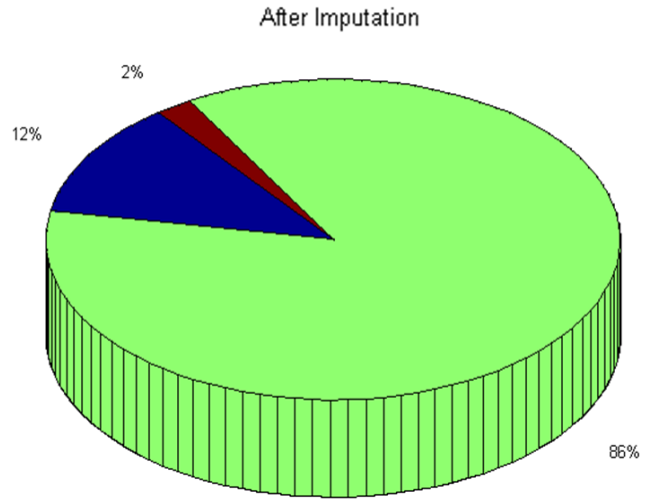

Figure 3. Classified and Misclassified Structure of Dataset After Imputation

The above figure 2 & 3 Shows the structure of dataset to predict classified and misclassified details of customer subscribe a term deposit. The above figure 2 indicate that, the structure of training dataset contains 88% of customer do not subscribe long term deposit, and 12% of customer do subscribe a term deposit in the bank. While imputing the data randomly using computation of cognitive approach applying pattern knowledge theory, could find changes in the dataset. Figure 3 indicate 2% of customer predicts to subscribe a term deposit from the misclassified value 'No'

V. RESULT AND DISCUSSION

Following Table 1 describes the categorization of various prediction models using experimental analysis in terms of accuracy rate. Categorization of prediction model shows that, the customer subscribe a long term deposit or not, which representing the term namely subscriber "No", Misclassification of "No", Subscriber "Yes" and Misclassification of "Yes". The following category analysing the accuracy rate using supervised prediction such as Decision Tree (DT), Naïve Bayes (NB), AdaBoosting (ADAB) , Bolzano, and COBACO models. Compare to all the positive accuracy rate of prediction, COBACO model produce high accuracy results.

EXPERIMENTAL ANALYSIS IN TERMS OF SUBSCRIBE ACCURACY RATE					
Categorization	Decision Tree	Naïve Bayes	AdaBoosting	Bolzano	COBACO
Subscriber "No"	93.66%	92.50%	93.33%	93.16%	92.48%
Misclassification "No"	6.34%	7.49%	6.67%	6.84%	7.52%
Subscriber "Yes"	52.15%	64.40%	45.15%	52.55%	65.10%
Misclassification "Yes	47.85%	35.60%	54.85%	47.45%	34.90%

Table 1 Summary of Accuracy Results for Prediction Models

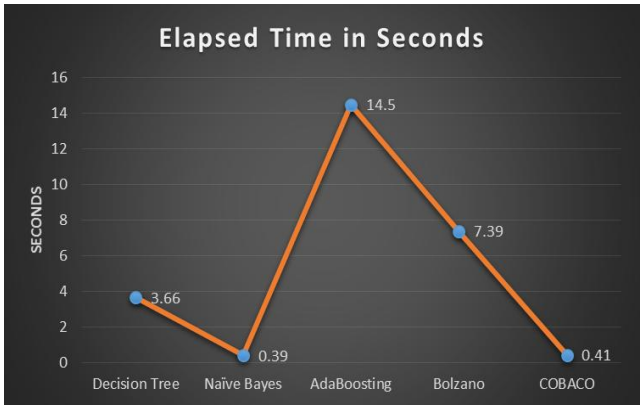

Figure 4 Comparison of Elapsed Time for Prediction Models

Above figure 4 describes the relation between various prediction models of Elapsed time calculation in seconds. Categorization of prediction model shows that, the customer subscribe a long term deposit or not, which represent the term namely subscriber "No", Misclassification of Subscriber "No", Subscriber "Yes" and Misclassification of "Yes". The following comparison state that, the Elapsed time for various prediction models such as Decision Tree (DT), Naïve Bayes (NB), AdaBoosting (ADAB) , Bolzano, and COBACO.

Various prediction model analysis of machine learning techniques such as Naïve Bayesian, Decision Tree, Adaptive Boosting, implementing mathematical approach like Bolzano Weierstrass Theorem, and COBACO model. Analyzed the COBACO model using Cognitive Pattern Knowledge Theory in order to predict misclassified value of subscriber in both the categories as "Misclassification of Subscriber Yes" and

"Misclassification of Subscriber No". COBACO produce high accuracy rate calculation than the other model. Compare to all models Adaptive Boosting approach elapsed time higher than the other model. Naïve Bayesian model performed in minimum elapsed time; hence, Pattern Knowledge Approach is unadoptable. But COBACO model capable of adapting Cognitive Pattern Knowledge Theory to predict the misclassification rate of hidden attributes in large dataset. The result of this experiment analysis shows that COBACO model more substantial than the other model.

VI. CONCLUSION

Machine learning techniques propose wonderful approaches for building predictive models from large dataset. Prediction of misclassification data arises common issues in machine learning analysis. Hence by applying such approaches in usual procedure need to overcome various challenges such as not executable to access the data centralized due to the massive size of the data sets, limitation to access the data sources, and produce bias results while accessing incomplete data. This thesis focuses on the problem of missing data and addressed five aspects of predictive modeling they are missing data problem, various predictive approaches in machine learning techniques, introduced effective Cognitive Bayes Computation (COBACO) approaches, compared all predictive techniques for the selection of predictive techniques.

FUTURE WORK

Future research topics initiate from current study indicates some promising directions include two possible research topics are presented as part of this thesis. First, predictive modeling of misclassification data using COBACO may be expanded with new concept of research may include the use of COBACO with larger datasets than the existing research.

Second research topic will be reducing the computation or elapsed time of COBACO approach. From the existing research work, the experimental results generate the elapsed time for all the approaches. Compare to all prediction approach Naïve Bayesian take less elapsed time than COBACO, the reason behind, COBACO generate all attributes in the dataset using pattern knowledge theory. In case of Naïve Bayesian techniques, unable to apply cognitive pattern theory. In such case COBACO approach is more effective and perform efficient performance than other techniques. Applications of the proposed classifier design for making predictions in real time, can predict the probability of multiple classes of the target variable, also perform text classification with a higher success rate, spam filtering and identify positive and negative sentiment analysis

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* 39 (2012) 2590-2596.
2. Gayan Prasad Hettiarachchi, Dhammika Suresh Hettiarachchi, NadeekaNilminiHettiarachchi, and Azusa Ebisuya, "Next Generation Data Classification and Linkage" *IEEE Global Humanitarian Technology Conference* 2014
3. Pei Zhang, Xiaoyu Wu, Xiaojun Wang, and Sheng Bi, "Short-Term Load Forecasting Based on Big Data Technologies" *CSEE Journal of Power and Energy Systems*, Vol. 1, No. 3, September 2015.
4. Ryo Karakida, Masato Okada, and Shun-Ichi Amari, "Dynamical Analysis of Contrastive Divergence Learning: Restricted Boltzmann Machines with Gaussian Visible Units" *Neural Networks* 79 (2016) 78-87, Journal homepage: www.elsevier.com/locate/neunet
5. Seema Sharma, Jitendra Agrawal, and Sanjeev Sharma, "Classification through Machine Learning Technique: C4.5 Algorithm Based on Various Entropies" *International Journal of Computer Applications* (0975-8887), Vol. 82, No. 16, November 2013.
6. https://en.wikipedia.org/wiki/Machine_learning
7. Yan Li, Manoj A. Thomas, "A Multiple Criteria Decision Analysis (MCDA) Software Selection Framework" *47th Hawaii International Conference on Systems Science* 2014
8. 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* Vol. 2, 2014
9. Wei Huang, "A Novel Disease Severity Prediction Scheme via Big Pair-Wise Ranking and Learning Techniques Using Image-Based Personal Clinical Data" *Signal Processing* 124 (2016) 233-245 Journal homepage: www.elsevier.com/locate/sigpro.
10. AlirezaFarhangfar, Lukasz Kurgan, and WitoldPedrycz, "Experimental Analysis of Methods for Imputation of Missing Values in Databases" *proceedings of the SPIE*, Vol. 5421, pp172-182, 04/2004
11. Edgar Acuna, and Caroline Rodriguez, "The Treatment of Missing Values and its Effect in the Classifier Accuracy" *Proceedings of the Meeting of the International Federation of Classification Societies* (IFCS) ISSN 1431-8814, [Online] link.springer.com/chapter pp639-647, 2004
12. Peng Liu, and Lei Lei "Missing Data Treatment Methods and NBI Model" *Proceedings of Sixth International Conference on Intelligent Systems Design and Applications* (ISDA'06)
13. Kai Zhang, Liang Lan, James T. Kwok, Slobodan Vucetic, and BahramParvin, "Scaling Up Graph-Based Semisupervised Learning via Prototype Vector Machines" *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 3, March 2015
14. MiroslawPawlak "Kernel Classification Rules from Missing Data" *IEEE Transactions on Information Theory*, Vol. 39, No.3, May 1993

AUTHORS PROFILE

Dr. S.Kanchana received PhD from Bharathiyar University, India in 2018 in the area of Data Mining. Currently she is working as Assistant Professor, SRM Institute of Science and Technology, Chennai, India. She has 12years of teaching experience. She has published 15 technical papers in various reputed International Journals. She attended more than 5 International / National conferences. Her areas of interests include Machine Learning, Deep Learning and semantic web. She is Life Member of ISCA.