"Missing Data Classification In Data Mining Applications For Network Performance Improvement Using Artificial Intelligence"

Gopal Patil (Ph.D. Research Scholar)[1]

Department of Computer Science and Application

Sarvepalli Radhakrishnan University

NH 12, RKDF IST CAMPUS, HOSHANGABAD ROAD, MISROD, BHOPAL (M.P.)

Dr. Raj Thaneeghaivel.V[2] (Assistant Professor)

Department of Computer Science and Application

Sarvepalli Radhakrishnan University

NH 12, RKDF IST CAMPUS, HOSHANGABAD ROAD, MISROD, BHOPAL (M.P.)

Abstract

In the current year number of organization used the IOT device for basic operation, such as medical,hospital, smart city,industry 4.0, etc. A variety of studies have therefore been performed on specific technology, such as context recognition, vice suggestion and missing data classification in IoT applications, using data created from data mining applications. These studies performance improvement that the data produced by data mining applications is complete. Many IoT device generated missing data just like missing value, immoral value Noisy values. We proposed technique missing data classification using Cross Neural Network(CNN).we represent through experiment our proposed approach very effective for Missing data classification.

Keyword: Cross Neural Network , IoT applications , Internet of Things , Missing values

Introduction

In this respect, when we were analysing the dataset obtained from a smart space with several IoT sensors, we noticed a persistent missing pattern that was very distinct from the missing data value. missing of successive missed values for a few seconds and up to a few hours.The pattern is thus a critical factor in the availability and efficiency of IoT applications; however it cannot be overcome by current imputation methods with a missing value. A novel approach to the imputation of missed value missing pattern is therefore necessary. the imputation of missing values is possible by the learning of other data streams associated with this data stream. The experimental findings reveal that our proposed solution boosts imputation efficiency. Our solution may therefore be a promising technique that enables IoT applications[1]And facilities with a good lack-value imputation accuracy. Miss-value imputation is a serious and difficult concern in the IoT world. A smart space with a multi-IoT-device ecosystem was considered. For up to a few hours, the persistent missing-value The series that was never dealt with before involves blocks over a few seconds of successive missing-values.

To calculate the missing values of the continuous missing pattern, we discovered the idea of the connection between the IoT data streams. we proposed a deep learning model called Cross Neural Networkfor the non-value imputation of the continuous missing patternwhich uses several dedicated ANN for the creation of features and a completely connected layer.

Using a correlation-based framework, ANNare preserves input data size to the incredibly long data sources, but can also cope with constant missed-value trends and high missing rates of heterogeneous data are used to mant devices. We tested Cross Neural Network on a real IoT dataset, to the results findings revealed that the proposed Cross Neural Network model greatly enhances imputation accuracy relative to traditional imputation algorithms. In addition, we have checked the value of the data association on input output by experiments.

Related work

Most data collection methods, such as survey surveys, field experiments, experimental research findings, etc., yield large quantities of knowledge. Missing values are inevitable in the data gathered. In addition, data mining methods, such as clustering and classification, have been developed to explore and uncover information from data that is complete, i.e. does not contain missing values. The existence of missing values limits the efficiency of these data analysis strategies. Methods to enhance the data mining process have also become a field that has interested many researchers. Generally, the researcher has two choices to build a dataset that does not have any missed values. I deleting or ignoring these inaccurate records with missing values (ii) filling the missing value with approximate values The deleting or ignoring rows with missing values has been found to be unreliable and thus more focus has been given to methods that forecast missing values. Some of these approaches relevant to the latest research work are discussed in this chapter. As the present research work focus on classification of data with missing data, the first section starts with a study on the works related to general classification, followed by various studies that have focused on missing values.

Raji, et al[1]This paper develops a real-time recording of the patient's vital signs using wearable monitors. Without the aid of the nurse, the patient will know the vital signs from the sensors and the device will store the sensor value in the form of a text document. The device is trained for vital sign data by using data mining approaches. Patients upload a text document to the system where in fact, they know their health status without any support from a nurse. This method allows high-risk patients to be tested in a timely manner and increase the quality of life of patients.

I. Mehmood et al[2]Produced a facial picture dataset, involving single and aggregate facial pictures. This information assortment can be utilized by different specialists as a benchmark for examination with other facial picture recovery frameworks progressively. Trial discoveries show that our proposed gadget outflanks other cutting edge strategies for extraction of highlights as far as execution and recuperation of IoT-helped energy-obliged stages.

S. Karimi-Bidhendi et al[3]Then a multilayer perceptron (MLP) with three hidden layers and a softmax output activation feature is used to obtain the final classification. Our approach provides favourable outcomes for UTS preparation and forecasting while providing superior results for MTS datasets. Unlike several published results, our approach is stable in the case of variable time series with missed data points and scales well with dataset sizes.

G. D. Kalyankar et al[4]Predictive analysis is a tool that incorporates a variety of data mining methods, machine learning algorithms and analytics to use present and historical data sets to gain insight and forecast potential threats.

Johnson I. Agbinya et al[5], the creator discarded the inconsistencies frequently connected with these standards. The instructional part of the book and the applications submitted are a portion of the reasons why the book is ideal for undergrad, postgraduate and enormous information examination aficionados. This content ought to lighten the fear of science frequently connected with reasonable information handling and advance quick uses of man-made brainpower, ecological sensor information displaying and examination, wellbeing data innovation.

A. Walinjkar et al[6] A very critical feature of customised health care is to constantly monitor the health of the client using wearable biomedical instruments and to analyse and if possible, anticipate future health risks that could prove lethal if not treated in a timely manner. The predictive feature of the method helps to prevent complications in delivering prompt medical care, often before a person enters a serious situation. Given the existence of state-of-the-art wearable health tracking systems, these devices tend to lack real-time data and predictive elements.

MISSING DATA HANDLING METHODS

Inspite of the various classification algorithms present, when presented with incomplete datasets, as mentioned previously, the accuracy of the classifier decreases. In these situations, special methods have to be developed to fill the missing values efficiently so as to maintain the accuracy of the classification process. Some of these methods are discussed in this section.

Imputation-based studies

A comparison of the different types of imputation. Methods that do not include imputation have also been suggested. These approaches have the advantage of using a cost-sensitive learning tool. It provides an excellent study on the various forms of cost-sensitive learning expenses, the most important of which are misclassification costs and testing costs. Most analysis has been conducted in recent years on non-uniform misclassification (only) costs, such as some previous work, such as considering the research cost alone without adding misclassification costs, which is evidently an oversight. Some researchers recognise both misclassification and test costs, but their approaches are less computationally effective since our methodology is focused on decision trees.

Multiple Imputation:

The principle of multiple imputations, first suggested, is to measure more than one value for the missing object. The benefit of multiple imputation is that it reflects the ambiguity of the attribute to be imputed. This is in contrast to the imputation of the mean answer, which does not contain the degree of ambiguity as to the meaning to be imputed. Therefore, analyses that consider imputed values much as observable values usually neglect uncertainty. Multiple imputation may be done either for longitudinal measurements or for a single response. The general technique for multiple imputation is to substitute and missing value with two or three values from the acceptable distribution of the missing values. This provides two or three total data sets. Repeated draws are taken from the post-predictive distribution of missed values. . As in fact, implicit models should be used instead of explicit models. [9] Address propensity-based imputation where one model is the likelihood of staying in the sample according to the covariate vector observed.

Missing data Classification is the method of discovering a collection of models (or functions) that represent. The derived model is based on the interpretation of the training data collection. The derived model can be described in a number of ways, such as classification (IF-THEN) laws, CNN, formulas or neural networks. Out of these, the use decision tress representation is more common as it can be quickly translated to classification rules. Classification can be used to predict the data objects mark class. However, in many implementations, users may choose to forecast any lost or inaccessible data values rather than class labels. This is typically the case because the expected values are numerical data and are often directly referred to as forecasts. While prediction can refer to both data value Missing data and class mark Missing data , it is generally limited to data value Missing data and is therefore distinct from classification. Prediction also involves the detection of patterns in distribution on the basis of available evidence.

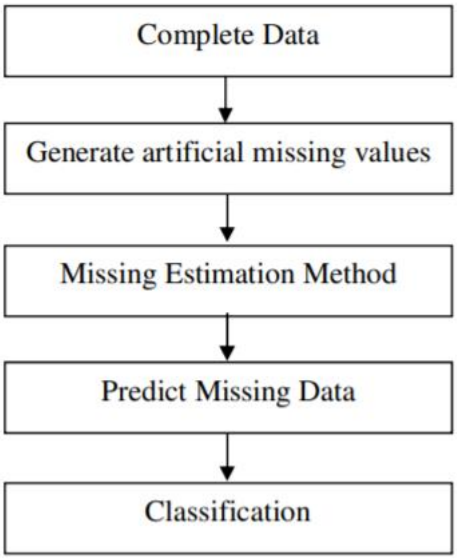

Figure 1: missing data classification process

For this purpose, the research study was divided into four phases of the study consists of two steps. 1. Propose methods that enhance imputation and non-imputation methods 2. Compare the performance of imputation and non-imputation based techniques for handling missing data for classification and compare the performance of the proposed systems with Mean and Median Imputation, KNN Imputation and Hot Deck Imputation Both the steps judge both missing value and classification performance. The imputation method proposed by is enhanced. The proposed non-imputation method enhances the model proposed . Both the models have been proved to be efficient in terms of missing value imputation and classification. However, when provided with large datasets, the time taken to complete the task is high. This research work proposes a technique that can reduce this time complexity and which can be combined with the above two models in a simple fashion. The proposed models are referred to as deep learning algorithm.

of the study focus on using association rules for handling missing data and use and analyze this resultant database for classification. The main objective here is to develop a model that combines Cross Neural Network method for classification. The system is referred to as Cross Neural Network method and is performed in two steps. The first step imputes the basis of association rules by comparing the antecedent part of rules with the known attribute values of missing value observation. The second step is triggered for the case when there is no association rule fired against any missing value and the values are imputed using Cross Neural Network approach. While creating association rules a partial matching concept is used. This model is enhanced in the following manner. First Method : The crucial factor of this method is the calculation of support and confidence thresholds. Too low or too high values of support and confidence will have negative impact on classification results. Therefore, in this research the calculation of these two values are eliminated by using confidence-lift-based support constraint and collective support constraint techniques that automatically calculate support and confidence Second Method : Here, the support and confidence parameters are refined to express the missingness. The support parameter is modified to include only those tupels that have no missing values. A similar correction is made to the confidence also, that is, to include only tuples with no missing values. Apart from these two measures, two new measures, namely, representativity and extensibility are included. The representativity is defined as the number of tuples that have no missing values divided by the number of tuples in the database. An itemset X is called extensible, if an itemset Y exists; such that X ∪ Y is frequent and representative. Further during experimentation, it was noticed that both models produces best result for small and medium sized databases, but its performance degraded with large database. To solve this problem, a database partition-based algorithm is used. This algorithm divides the database into portions which can be fit into RAM easily.

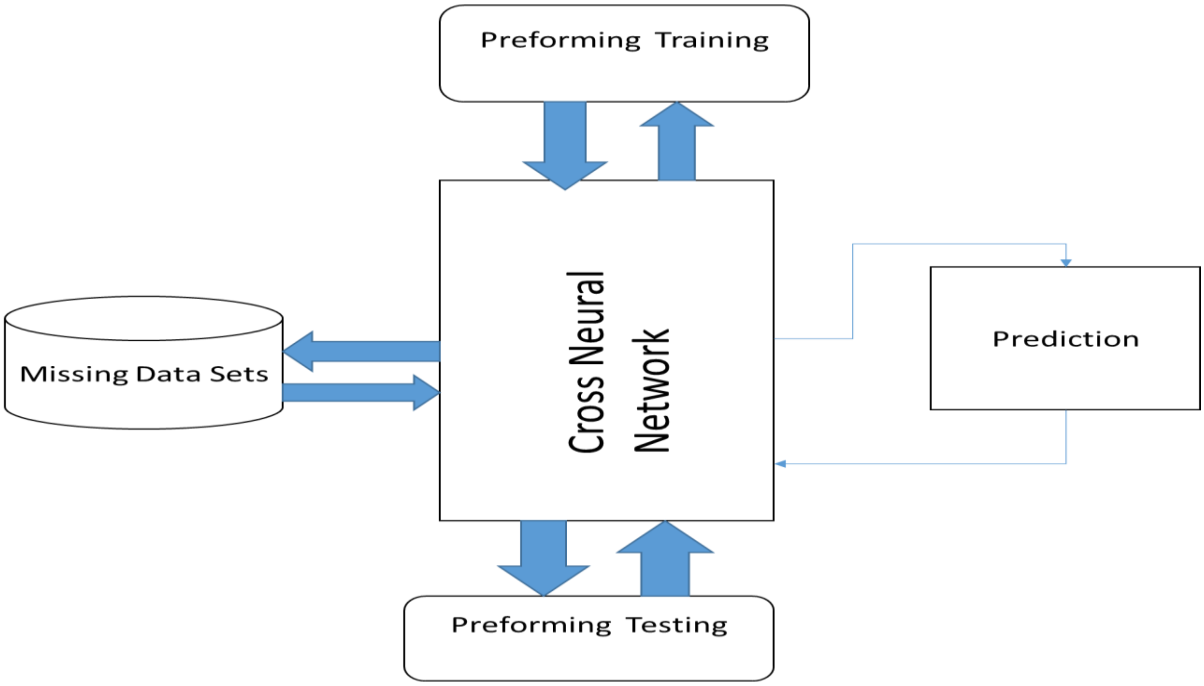

Figure2: Proposed model for missing data classification

PERFORMANCE METRICS

The fast development of interest in utilizing Cross Neural Network designs in different areas has been helped by the dispatch of many Cross Neural Network lately We have demonstrated that the technique works splendidly as far as low percent inclination, low mean square error and high inclusion. The re-enactments in this research explored the impact of the solicitation and the quantity of attributions based on a foreordained arrangement of conditions, the greater part of all, obliviousness. The discoveries show that neither the solicitation nor the measure of ascriptions significantly affect inclination, mean square error, or inclusion under these conditions.Be it as it might, this work offers a structure of trends for more complicated cases and more complex conclusions regarding missing values and the classification of missing data. The overall process of all proposed missing value handling procedures starts with introducing missing values in a complete dataset. After which, using various estimation methods the missing values are predicted. The impact of these predicted values on classification is then analyzed.

Many of the studies used the following performance metrics to test the proposed models.

Accuracy The accuracy of the designation is determined

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \text{ x } 100$$

Normalized Root Mean Square Error (NRMSE) The potency of missed values Imputation was calculated by the Normalized Root Mean Square Error (NRMSE) Calculated by

$$NRMSE = \frac{\sqrt{mean[(y_{true} - y_{imp})^2]}}{variance(y_{true})}$$

In this work we calculated the value of mean and variance that useful for missing data entry estimation.

y true represent the value of total data matrix and yimp represent as the input value of data matrix.

If the NRMSE turn out to be as a zero then getting results is correct. If the value reach low then NRMSE show the value of 1.

Compute the time used for execution : classified time used for missing value execution and complete the classification.

CONCLUSION

In recent years, Cross Neural Network and IoT have attracted the interest of analysts and industry verticals, many emerging developments have proved to have a positive influence on our lives, communities and the planet. Multi-stage multiple imputations are one tool for working with missing data that tracks the fluctuation of incomplete data. This technique, as such, is used by filling in conceivable values a few times to make a few final data sets and then correctly joining full data gauges using unique consolidation rules. . We introduced the latest novel technique and the related correlation of different classifiers needed for implementation. By way of simulations, we have shown that we have an effective estimator under suspicion. Our proposed approach very effective to classified the missing data.

Reference

[1]. Raji, P. G. Jeyasheeli and T. Jenitha, "IoT based classification of vital signs data for chronic disease monitoring," 2016 10th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2016, pp. 1-5, doi: 10.1109/ISCO.2016.7727048.

[2]. Mehmood et al., "Efficient Image Recognition and Retrieval on IoT-Assisted Energy-Constrained Platforms From Big Data Repositories," in IEEE Internet of Things Journal, vol. 6, no. 6, pp. 9246-9255, Dec. 2019, doi: 10.1109/JIOT.2019.2896151.

[3]. S. Karimi-Bidhendi, F. Munshi and A. Munshi, "Scalable Classification of Univariate and Multivariate Time Series," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 1598-1605, doi: 10.1109/BigData.2018.8621889.

[4]. G. D. Kalyankar, S. R. Poojara and N. V. Dharwadkar, "Predictive analysis of diabetic patient data using machine learning and Hadoop," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, 2017, pp. 619-624, doi: 10.1109/I-SMAC.2017.8058253.

[5]. Johnson I. Agbinya, "17 Probabilistic Neural Network Classifiers for IoT Data Classification," in Applied Data Analytics – Principles and Applications , River Publishers, 2020, pp.277-290.

[6]. Walinjkar and J. Woods, "ECG classification and prognostic approach towards personalized healthcare," 2017 International Conference On Social Media, Wearable And Web Analytics (Social Media), London, 2017, pp. 1-8, doi: 10.1109/SOCIALMEDIA.2017.8057360.

[7]. X. Fafoutis and L. Marchegiani, "Rethinking IoT Network Reliability in the Era of Machine Learning," 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Atlanta, GA, USA, 2019, pp. 1112-1119, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00189.

[8]. S. Ge, Q. Ye, Z. Luo and S. Zhao, "Try Everything: Detecting Occluded Faces by Cascading Outrageous Proposal Generation and Deep Convolutional Neural Network," 2017 IEEE Third International Conference on Multimedia Big Data (BigMM), Laguna Hills, CA, 2017, pp. 193-196, doi: 10.1109/BigMM.2017.40.

[9]. Sartori F., Melen R., Giudici F. (2019) IoT Data Validation Using Spatial and Temporal Correlations. In: Garoufallou E., Fallucchi F., William De Luca E. (eds) Metadata and Semantic Research. MTSR 2019. Communications in Computer and Information Science, vol 1057. Springer, Cham. https://doi.org/10.1007/978-3-030-36599-8_7

[10]. Firouzi F., Farahani B., Ye F., Barzegari M. (2020) Machine Learning for IoT. In: Firouzi F., Chakrabarty K., Nassif S. (eds) Intelligent Internet of Things. Springer, Cham. https://doi.org/10.1007/978-3-030-30367-9_5

[11]. Rajawat A.S., Upadhyay P., Upadhyay A. (2021) Novel Deep Learning Model for Uncertainty Prediction in Mobile Computing. In: Arai K., Kapoor S., Bhatia R. (eds) Intelligent Systems and Applications. IntelliSys 2020. Advances in Intelligent Systems and Computing, vol 1250. Springer, Cham. https://doi.org/10.1007/978-3-030-55180-3_49

[12]. Yen, N.Y., Chang, J., Liao, J. *et al.* Analysis of interpolation algorithms for the missing values in IoT time series: a case of air quality in Taiwan. *J Supercomput* **76,** 6475–6500 (2020). https://doi.org/10.1007/s11227-019-02991-7

[13]. Hiriyannaiah S., Khan Z., Singh A., Siddesh G.M., Srinivasa K.G. (2020) Data Reduction Techniques in Fog Data Analytics for IoT Applications. In: Tanwar S. (eds) Fog Data Analytics for IoT Applications. Studies in Big Data, vol 76. Springer, Singapore. https://doi.org/10.1007/978-981-15-6044-6_12

[14]. Fisher, P.S., James, J., Baek, J. *et al.* Mining intelligent solution to compensate missing data context of medical IoT devices. *Pers Ubiquit Comput* **22,** 219–224 (2018). https://doi.org/10.1007/s00779-017-1106-1

[15]. M. Mohammadi, A. Al-Fuqaha, S. Sorour and M. Guizani, "Deep Learning for IoT Big Data and Streaming Analytics: A Survey," in IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 2923-2960, Fourthquarter 2018, doi: 10.1109/COMST.2018.2844341.

[16]. A. Singh Rajawat and S. Jain, "Fusion Deep Learning Based on Back Propagation Neural Network for Personalization," 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, 2020, pp. 1-7, doi: 10.1109/IDEA49133.2020.9170693.

[17]. Ozan E.C., Riabchenko E., Kiranyaz S., Gabbouj M. (2016) An Optimized k-NN Approach for Classification on Imbalanced Datasets with Missing Data. In: Boström H., Knobbe A., Soares C., Papapetrou P. (eds) Advances in Intelligent Data Analysis XV. IDA 2016. Lecture Notes in Computer Science, vol 9897. Springer, Cham. https://doi.org/10.1007/978-3-319-46349-0_34

[18]. Abdelkhalek I., Ben Brahim A., Essousi N. (2018) A New Way of Handling Missing Data in Multi-source Classification Based on Adaptive Imputation. In: Abdelwahed E., Bellatreche L., Golfarelli M., Méry D., Ordonez C. (eds) Model and Data Engineering. MEDI 2018. Lecture Notes in Computer Science, vol 11163. Springer, Cham. https://doi.org/10.1007/978-3-030-00856-7_8

[19]. Porro-Muñoz D., Duin R.P.W., Talavera I. (2013) Missing Values in Dissimilarity-Based Classification of Multi-way Data. In: Ruiz-Shulcloper J., Sanniti di Baja G. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2013. Lecture Notes in Computer Science, vol 8258. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41822-8_27

[20]. Nguyen, T.T., Tsoy, Y. A kernel PLS based classification method with missing data handling. *Stat Papers* **58,** 211–225 (2017). https://doi.org/10.1007/s00362-015-0694-y

[21]. Salleh M.N.M., Samat N.A. (2017) An Imputation for Missing Data Features Based on Fuzzy Swarm Approach in Heart Disease Classification. In: Tan Y., Takagi H., Shi Y., Niu B. (eds) Advances in Swarm Intelligence. ICSI 2017. Lecture Notes in Computer Science, vol 10386. Springer, Cham. https://doi.org/10.1007/978-3-319-61833-3_30

[22]. Lagona F., Picone M. (2013) Classification of Multivariate Linear-Circular Data with Nonignorable Missing Values. In: Grigoletto M., Lisi F., Petrone S. (eds) Complex Models and Computational Methods in Statistics. Contributions to Statistics. Springer, Milano. https://doi.org/10.1007/978-88-470-2871-5_13