Soft Computing Technique Based on Missing Value Treatment

Sukhman Kaur

Abstract: Missing value treatment is an actual yet challenging issue confronted in data mining. In existing work missing value treatment is a procedure that replaces the missing values in a dataset by some conceivable values. The conceivable values are generally generated from the dataset using a statistical evaluation. These types of results don’t give accurate outcomes. In this paper, soft computing is used in the random forest approach using for missing value treatment that is devised and implemented on the different types of social media. Using random forest approach results are improved from existing techniques.

Index Terms: Missing value, Missing value treatment, Soft computing, Social media.

I. INTRODUCTION

The development of online networking throughout the most recent decade has reformed the manner in which people collaborate and ventures direct business. People produce information at a remarkable rate by connecting, sharing, and devouring substance through web-based social networking. Comprehension and handling this new sort of information to gather noteworthy examples presents difficulties and open doors for interdisciplinary research and apparatus advancement [4]. Internet-based life Mining incorporates web-based life, informal organization examination, and information mining to give an advantageous and sound stage for understudies, specialists, analysts, and venture administrators to comprehend the nuts and bolts and possibilities of web-based life mining [9]. Missing values are a common problem in most clear research spaces, for example, Social Media Analysis, Satellite Data, GPS Data, Biology, Medicine or Climatic Science. They can ascend out of various sources, for example, abusing of tests, low standard to-change degree, estimation goof, non-reaction or killed intriguing esteem. Missing values make it troublesome for investigators to perform information examination. Three kinds of issues are typically connected with missingworth:

(1) Loss of efficiency.
(2) Complications in handling and analyzing the data.
(3) Bias resulting from differences between missing and complete data.

Statistician categorized missing data into three categories:

- Missing not at Random (MNAR).
- Missing at Random (MAR).
- Missing completely at Random (MCAR).

The following are common methods:

* Hot deck: A randomly picked a value from an individual who has comparable values on different factors.
* Cold deck: An efficiently picked a value from an individual who has comparable values on a different factor.
* Regression: The forecasted value obtained by regression the missing variable on different factors.
* Stochastic regression: The forecasted value from a regression plus an arbitrary residual value.
* Interpolation and extrapolation: An expected value from different perceptions from a similar individual.

Fengfeng Fan, 2017 presents the work on internet missing worth attribution utilizing OL-MVI Model. In this methodology, all the records are broke down and outlined for missing value treatment from constant information. Davis et al. [14] underline the assessment of different internet-based life with investigation designs. The online life incorporates Google+, LinkedIn, Facebook and Twitter with their relative ways to deal with brings the information and reconciliation of social profiles. What's more, the standard digging joining is introduced for various applications by the creators. Barve A. et al. [25] 2018 SVN, KNN, and Random Forest Approach worked on the dataset of social media. The anticipated outcomes are displayed to be efficacious and execution mindful on the particular datasets.

II. PROBLEM DEFINITION AND RELATED WORK

Identification of the problem is that in the statistical approach and algorithms to deal with the missing value include the average value and removal of the records from the real dataset. If such a strategy is used, the results cannot be evaluated. In existing research work of missing value treatment, the statistical evaluation using mean, linear regression of existing values are used which can be improved using soft computing algorithm.

Statistical Based Treatment is having an evident focus on the mean and prediction based on that curve is done on scenarios of regression and prediction based treatment.

III. MISSING VALUE TREATMENT USING SOFT COMPUTING

This work is having the key focus on the randomly based imputation approach is having the fitness score based on the final outcome and overall acceptability score.
values. These missing values are exclusively prepared utilizing the approach is discretionary choice trees utilizing random forest paradigm. The choice tree in each gathering is handled to have the area mean occurrences and hence the outcomes or result of expected mean values is assessed from each set. The ascribed an incentive from every choice tree is related with a particular acknowledgment score of positioning and this kind of scoring is utilized to at long last have the last credited esteem which is the best fit for attribution.

IV. ALGORITHM IMPUTATION OF MISSING VALUE IN THE FETCHED DATASET

PHASE 1: Using Association Rule Mining for Imputation.
1) Activation of Rule Mining and Extraction of LiveData.
2) Generate rules and distance based prediction from the training dataset.
3) From the list of association rules remove rules with consequent containing value “Missing or Null”.
4) Sort association rules by confidence in descending order and apply soft computing for Fitness of the Imputation and Best Fit Candidate for Missing Value.
5) If suitable association rule was found, fill the missing value in the data set for missing values imputation by value in the consequent of the association rule.

PHASE 2: Clustering if there were no suitable association rule.
1) In the preparation informational index fill every single missing an incentive by the uncommon value. In further advances are these values called "Missing or Null". Apply Density-based Clustering from the datasets of filled values.
2) Produce association rules from the preparation informational dataset.
3) From the list of association rules evacuate rules with help lower than required.
4) From the list of association rules remove rules with consequent that is combination longer than a threshold value.
5) From the list of association rules expel rules with resulting containing the value "Missing or Null".
6) Sort association rules by trust in the diving request.
7) If suitable association rule was found, fill missing value in the data set for missing values imputation by value in consequent of the association rule. Else fill missing value in the data set for missing values treatment by the most common attribute value (except the value "Missing or Null").

In the above mentioned algorithm, there are two phases. The first phase deals with the data extraction and pre-processing so that the further evaluation of missing values and imputation can be done. These second phases having focus on the treatment of missing values using soft computing approach in integration with random forest algorithm. Random Forest algorithm is a prominent algorithm for the optimization and population based processing of datasets.

Flow of Work Flow Diagram

V. IMPLEMENTATION AND EXPERIMENTAL RESULT

Python Programming is utilized for actualizing the irregular Random Forest Approach and information is brought from various web-based lives utilizing APIs. In this examination deal with missing value treatment, the circumstance of expelling specific substance is taken so the client conviction related to that word can be evaluated with missing value treatment. The substance or customer feelings in light of different classes are taken with the live extraction of data from different online interpersonal interaction. These classes are used so the all-out decision about the zone, guidance, and district can be peniless down with the extraction of data related to customer courses of occasions. In most of the tweets, the evaluation or customer feelings of electronic interpersonal interaction are related to the academicians and specialists who show their points of view or ends on this online life. The assessment of supporters is done in light of the fact that it is numeric esteem and distinguishing proof should be possible where there are unimportant or invalid adherents. In the event that we don't think about these values, the outcomes can be inefficient to the last end. In customary system, such characteristics based records are deleted anyway by this methodology a definitive outcome cannot be extraordinary. In the isolated tweets and messages, it is found that the investigation perspectives and standards set around different classes are
generally discussed by the examiners, academicians and government officials. With the execution of proposed approach, the missing values ascertainment is seen to be solid with the filling or attribution of the characteristics using an irregular woodland approach that is one the indisputable methodology settling on the use of huge decision trees and a short time later last situating of the best outcome with the scoring of results.

A. RAW DATA AFTER PREPROCESSING WITH MISSING VALUE

Following screenshot resent that the data is collected from the different social media and processed it. It shows data into CSV (comma-separated values) format.

Table I DATA WITH MISSING VALUE

| UserID | Followers | Created Date | Created Month | Created Year |
|--------|-----------|--------------|---------------|--------------|
| 64376190 | 345861 | 6 | 4 | 2015 |
| 65020179 | 3164 | 20 | 12 | 2017 |
| 40678590 | 209755 | 8 | 2015 |
| 6341527 | 230564 | 1 | 11 | 2012 |
| 18938647 | MISSING / NULL | 3 | 10 | 2013 |
| 2297215 | MISSING / NULL | 23 | 9 | 2013 |
| 28174243 | 453932 | 11 | 6 | 2016 |
| 292039 | 51130 | 15 | 2 | 2017 |
| 70192405 | 287214 | 26 | 8 | 2017 |
| 23344003 | 225067 | 6 | 12 | 2017 |
| 129727219 | MISSING / NULL | 18 | 7 | 2013 |
| 10933427 | 453250 | 13 | 7 | 2017 |
| 4271124 | 87413 | 11 | 12 | 2016 |
| 8470559 | 58529 | 20 | 7 | 2014 |
| 5135929 | 240988 | 12 | 5 | 2014 |
| 30370636 | 35284 | 10 | 10 | 2014 |
| 34114564 | 65328 | 21 | 4 | 2012 |
| 70537456 | 370860 | 23 | 5 | 2016 |
| 59983505 | 460759 | 16 | 7 | 2017 |
| 18095739 | 395564 | 6 | 4 | 2015 |
| 5210524 | 478839 | 28 | 12 | 2016 |

B. MISSING VALUE IMPUTATION USING STATISTICAL VALUED BASED APPROACH

Following is the point of view of execution situation in programming language for missing value treatment.

Table II DATA WITH MISSING VALUE IMPUTATION USING STATISTICAL VALUED BASED APPROACH

| UserID | Followers | Created Date | Created Month | Created Year |
|--------|-----------|--------------|---------------|--------------|
| 64376190 | 345861 | 6 | 4 | 2015 |
| 65020179 | 3164 | 20 | 12 | 2017 |
| 40678590 | 209755 | 8 | 2015 |
| 6341527 | 230564 | 1 | 11 | 2012 |
| 18938647 | MISSING / NULL | 3 | 10 | 2013 |
| 2297215 | MISSING / NULL | 23 | 9 | 2013 |
| 28174243 | 453932 | 11 | 6 | 2016 |
| 292039 | 51130 | 15 | 2 | 2017 |
| 70192405 | 287214 | 26 | 8 | 2017 |
| 23344003 | 225067 | 6 | 12 | 2017 |
| 129727219 | MISSING / NULL | 18 | 7 | 2013 |
| 10933427 | 453250 | 13 | 7 | 2017 |
| 4271124 | 87413 | 11 | 12 | 2016 |
| 8470559 | 58529 | 20 | 7 | 2014 |
| 5135929 | 240988 | 12 | 5 | 2014 |
| 30370636 | 35284 | 10 | 10 | 2014 |
| 34114564 | 65328 | 21 | 4 | 2012 |
| 70537456 | 370860 | 23 | 5 | 2016 |
| 59983505 | 460759 | 16 | 7 | 2017 |
| 18095739 | 395564 | 6 | 4 | 2015 |
| 5210524 | 478839 | 28 | 12 | 2016 |

C. MISSING VALUE TREATMENT USING RANDOM BASED APPROACH

The screenshot is showing the treated values using the random forest-based approach of missing value treatment. These imputed values are different from the statistically based imputation.

Table III DATA WITH MISSING VALUE TREATMENT

Summary of the Records Identified with Missing Values Extracted from different social media

Following is the analytics of the general executions in exceptional social media in terms of lacking values extracted in unique timelines.

Following are the styles of data extracted from social media from distinct class and instance with the datasets used on assorted subjects.

Table IV DATASETS EVALUATION UNDER CLASSES AND INSTANCES

| Datasets | Instances | Dataset Size (KB) | Classes |
|----------|-----------|-------------------|---------|
| UGC      | 548       | 96                | 1       |
| Delhi    | 680       | 37                | 2       |
| Disaster | 697       | 48                | 2       |
| Election | 536       | 60                | 3       |
| University | 673   | 60                | 1       |
| Mumbai   | 905       | 46                | 2       |

Table V is the depiction of class definitions with the facts evaluation scenarios. Inside the subsequent effects, the accuracy degree is measured.
The accuracy is evaluated with the benchmark of traditional suggest. The evaluation of blunders aspect from the proposed method is evaluated and compared from statistical analysis. The deviation among the values obtained from the proposed and classical method is the bottom of the assessment of mistakes factor in absolute and relative terms.

Table V CLASS DEFINITIONS AND RECORDS

| Class    | Id | Records Evaluated |
|----------|----|-------------------|
| Education| 1  | 1219              |
| Location | 2  | 2211              |
| Politics | 3  | 487               |

Following table VI is the percentage of accuracy finished after implementation using specific processes of the existing

Table VI APPROACH BASED ACCURACY

| Category | Statistical Evaluation Based Outcome | Random Forest |
|----------|-------------------------------------|---------------|
| Education| 50                                  | 70            |
| Location | 66.48                               | 94            |
| Politics | 75.83                               | 95            |

D. EVALUATION AND COMPARISON OF INTEGRITY IN PROPOSED APPROACH AND EARLIER APPROACH

Integrity signifies the consistency of the set of rules in terms of jogging in distinctive key phrases. From the effects, it’s far obtusive that the proposed random forest method is integrity and consistency aware of exclusive situations of execution with special keywords in comparison to the preceding processes of statistical evaluation. Following the effects show that the execution time is integrity aware and constant without any ambiguity.

VI CONCLUSION

Statistical Evaluation Based techniques are used from many years so in this research work soft computing based technique is used. Using soft computing and Random Forest based algorithms the overall result can be improved. This work extracts the live data from multiple social and extracts the missing values. On extracted missing values, the global fitness score for missing value treatment is done for higher accuracy using soft computing. There are nature inspired approaches which can be further analyzed.

REFERENCES

1. Bifet, A., & Frank, E. Sentiment knowledge discovery in Twitter streaming data. In International Conference on Discovery Science. Springer Berlin Heidelberg, 2010.
2. Bollen, J., Mao, H., & Pepe, A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM, 11, 2009, 450-453.
3. Bollen, J., Mao, H., & Pepe, A. Determining the Public Mood State by Analysis of Microblogging Posts. In ALIFE2010. (pp. 667-668).
4. Asur, S., & Huberman, B. A. Predicting the future with social media. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on (Vol. 1, pp. 492-499). IEEE.
5. Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. User-level sentiment analysis incorporating social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1397-1405). ACM, 2011.
6. Saif, H., He, Y., & Alani, H. Semantic sentiment analysis of Twitter. In International Semantic Web Conference (pp.508-524), 2012. Springer Berlin Heidelberg.
7. Leong, C. K., Lee, Y. H., & Mak, W. K. Mining sentiments in SMS texts for teaching evaluation. Expert Systems with Applications, 39(3), 2584-2589, 2012.
8. Dong, H., Shahheidari, S., & Daud, M. N. R. B. Twitter sentiment mining: A multi-domain analysis. In Complex, Intelligent, and Software Intensive Systems (CISIS), 2013 Seventh International Conference (pp. 144-149). IEEE, 2013.
9. Cambria, E., Fu, J., Bisio, F., & Poria, S. AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis. In AAAI (pp. 508-514), 2013.
10. Al-Rubaiie H, Qiu R, Li D. Identifying Mulashwer software products through sentiment analysis of Arabic tweets. International Informatics and Computer Systems (CIICS), 2016 International Conference 2016 Mar 13 (pp. 1-6). IEEE, 2016.
11. Heredia B, Khoshgofaar TM, Prusa J, Crawford M. Cross-domain sentiment analysis: An empirical investigation. Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on 2016 Jul 28 (pp. 160-165). IEEE, 2016.
12. Blaz CC, Becker K. Sentiment analysis in tickets for its support. In Mining Software Repositories (MSR), 2016 IEEE/ACM 13th Working Conference on 2016 May 14 (pp. 235-246). IEEE, 2016.
13. Fiarni C, Maharani H, Pratama R. Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique. In Information and Communication Technology (ICoICT), 2016 4th International Conference on 2016 May 25 (pp. 1-6). IEEE, 2016.
14. lexicon-based sentiment analysis on Bahasa Indonesia. Engineering Seminar (InAES), International Annual 2016 Aug 1 (pp. 28-31). IEEE, 2016.
15. Nithya R, Maheswari D. Correlation of feature score to overall sentiment score for identifying the promising features. Computer Communication and Informatics (ICCICI), 2016 International Conference on 2016 Jan 7 (pp. 1-5). IEEE, 2016.
16. Bouazizi M, Ohtsuki TO. A pattern-Based approach for Sarcasm Detection on Twitter. IEEE Access. 2016;4:5477-88, 2016.
17. Gatti L, Guerini M, Turchi M. Sentirwors: Deriving highprecision and high coverage lexicon for sentiment analysis. IEEE Transactions on Affective Computing. 2016 Oct 1;7(4):409-21, 2016.
18. Biltawi M, Etaiwi W, Tedmori S, Hudaib A, Awajan A. Sentiment classification techniques for Arabic language: A survey. Information and Communication Systems (ICICS), 2016 7th International Conference on 2016 Apr 5 (pp. 339-346). IEEE,2016.

19. Rabab'ah AM, Ayyoub M, Jararweh Y, Al-Kabi MN. Evaluating sentistrength for arabic sentiment analysis. Computer Science and Information Technology (CSIT), 2016 7th International Conference on 2016 Jul 13 (pp. 1-6). IEEE,2016.

20. Barve, Abhishek, Manali Rahate, Ayesha Gaikwad, and Priyanka Patil. “Terror Attack Identifier: Classify using KNN, SVM, Random Forest algorithm and alert through messages.” International Research Journal of Engineering and Technology (IRJET) 2018.

21. Fengfeng Fan, 2017 presents the work on online missing value imputation using OL-MVI Model. In this approach, all the tuples / records are analyzed and summarized for missing value treatment from real timedata.

AUTHOR PROFILE

Sukhman Kaur received M.C.A degree in computer science and application from Punjabi University, Patiala, India, in 2015. She received an MPhil degree in computer science and application from Punjabi University, Patiala, India, in 2018. She is currently pursuing a Ph.D. degree in Computer Science at Punjabi University, Patiala, India. Her research interests include Data mining and networking.