rticle Information Overview

Manuscript ID

symmetry-2244914

Status

Website online

DOI

10.3390/sym15040887

Publication Certificate

Banner

Download Banner (PDF)

Website Links

Abstract HTML version PDF version Manuscript

Article type

Article

Title

Median-KNN Regressor-SMOTE-Tomek Links for Handling Missing and Imbalanced Data in Air Quality Prediction

Journal

Symmetry

Volume

15

Issue

4

Section

Computer Science and Symmetry/Asymmetry

Special Issue

Machine Learning and Data Analysis

Abstract

The Air Quality Index (AQI) dataset contains information on measurements of pollutants and ambient air quality conditions at certain location that can be used to predict air quality. Unfortunately, this dataset often has many missing observations and imbalanced classes. Both of these problems can affect the performance of the prediction model. In particular, predictions for the minority class are very important because inaccurate predictions can be fatal or cause big losses. Moreover, the missing data may lead to biased results. This paper proposes the single imputation of the median and the multiple imputations of the k-Nearest Neighbor (KNN) regressor to handle missing values of less than or equal to 10% and more than 10%, respectively. At the same time, the SMOTE-Tomek Links address the imbalanced class. These proposed approaches to handle both issues are then used to assess the air quality prediction of the India AQI dataset using Naive Bayes (NB), KNN, and C4.5. The five treatments show that the proposed method of the Median-KNN regressor-SMOTE-Tomek Links is able to improve the performance of the India air quality prediction model. In other words, the proposed method succeeds in overcoming the problems of missing values and class imbalance.

Keywords

1

air quality; missing values; imbalanced data; median; KNN; SMOTE-Tomek Links

Author Information

Submitting Author

Yulia Resti

Corresponding Author

Yulia Resti

Author #1

Winoto Chandra

Affiliation

1. Doctoral Study Program, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Jl. Padang Selasa Bukit Besar, Palembang 30139, Sumatera Selatan, Indonesia

2. Department of Information System, Faculty of Computer Science, Universitas Bina Darma, Jl. Jenderal A. Yani No. 3, Palembang 30111, Sumatera Selatan, Indonesia

E-Mail

080136216221001@student.unsri.ac.id (co-author email has not been published))

Author #2

Bambang Suprihatin

Affiliation

3. Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Jl. Raya Palembang-Prabumulih, Km.32, Inderalaya 30062, Sumatera Selatan, Indonesia

E-Mail

bambangs@unsri.ac.id (co-author email has not been published))

Author #3

Yulia Resti

Affiliation

3. Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Sriwijaya, Jl. Raya Palembang-Prabumulih, Km.32, Inderalaya 30062, Sumatera Selatan, Indonesia
E-Mail
yulia_resti@mipa.unsri.ac.id (corresponding author email)
Manuscript Information
Received Date
28 February 2023
Revised Date
30 March 2023
Accepted Date
3 April 2023
Published Date
9 April 2023
Submission to First Decision (Days)
23
Submission to Publication (Days)
39
Round of Revision
1
Size of PDF
3875 KiB
Word Count
5769
Page Count
16
Figure Count
10
Table Count
11
Reference Count
35
Editor Decision
Decision
Accept in current form
Decision Date
31 March 2023
Review Report
Reviewer 1
Review Report (Round 1)
Reviewer 2
Review Report (Round 1)
Reviewer 3
Review Report (Round 1)

Authors' Responses to Reviewer's Comments (Reviewer 1)

Author's Notes

Please see the attachment.

Author's Notes File

Report Notes

Review Report Form

Quality of English Language

( ) English very difficult to understand/incomprehensible

( ) Extensive editing of English language and style required

( ) Moderate English changes required

(x) English language and style are fine/minor spell check required

( ) I am not qualified to assess the quality of English in this paper

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	(x)	( )	( )	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	(x)	( )	( )	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	(x)	( )	( )	( )
Are the conclusions supported by the results?	(x)	( )	( )	( )

Comments and Suggestions for Authors

In this research, the author proposed the multiple imputations of the KNN regressor-SMOTE-Tomek links to handle the missing and imbalanced Air Quality Index data set, which can be used to predict air quality. This work shows a high potential for the improvement of the performance of the air quality prediction model.

There are some problems, which must be solved before it is considered for publication. If the following issues are well-addressed, this reviewer believes that the essential contribution of this paper is important for overcoming the problem of data missing and imbalance.

1. There are at least two grammatical and spelling errors in the manuscript, such as, in line18, one more full stop before “Moreover”; and in lines 230 and 233, “F-1 score’” should be “F1-score”; in line 306, one more full stop after “value of 100%“； in line 342, it should be “the more naive assumption…“; in line 386, it should be “using naive Bayes…“.
2. In Figure 3, the RMSE of k=22 seems lower than that of k=14, the author can give an inset to zoom up on this part. And in Figure 4, the lowest RMSE seems in k=3, the author can also give an inset in this figure.
3. The X and Y label seems too small in Figure 5, and readers may can not see them clearly.
4. In table 10, can the author demonstrate what the negative value represents, like the precision of LR DNA methylation?

5. Authors need to unify the format of all references, such as whether to add the month of the journal, like Ref 5, 17; whether to use both the starting and ending page numbers, like ref 7, 18; and whether to abbreviate the journal name.

Submission Date
28 February 2023
Date of this review
16 Mar 2023 07:27:14

Authors' Responses to Reviewer's Comments (Reviewer 2)
Author's Notes
Please see the attachment.
Author's Notes File
Report Notes
Review Report Form
Quality of English Language
( ) English very difficult to understand/incomprehensible
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I am not qualified to assess the quality of English in this paper

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	(x)	( )	( )	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	(x)	( )	( )	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	(x)	( )	( )	( )
Are the conclusions supported by the results?	(x)	( )	( )	( )

Comments and Suggestions for Authors
Dear Authors
overall mauscript is written very well, it is clear and right to the point. the quality of the Figures are low, please improve them and use white background for all of them.
Abstract:

Line 18: remove excessive “.”

Line 236-247: these are more of a methodology, please move them to method part. Here, you should focus on the results of the research.

Submission Date
28 February 2023
Date of this review
25 Mar 2023 09:14:50

Authors' Responses to Reviewer's Comments (Reviewer 3)
Author's Notes
Please see the attachment.
Author's Notes File
Report Notes
Review Report Form
Quality of English Language
( ) English very difficult to understand/incomprehensible
( ) Extensive editing of English language and style required
( ) Moderate English changes required
(x) English language and style are fine/minor spell check required
( ) I am not qualified to assess the quality of English in this paper

	Yes	Can be improved	Must be improved	Not applicable
Does the introduction provide sufficient background and include all relevant references?	(x)	( )	( )	( )
Are all the cited references relevant to the research?	(x)	( )	( )	( )
Is the research design appropriate?	(x)	( )	( )	( )
Are the methods adequately described?	(x)	( )	( )	( )
Are the results clearly presented?	(x)	( )	( )	( )
Are the conclusions supported by the results?	(x)	( )	( )	( )

Comments and Suggestions for Authors
The article is devoted to problem-solving (missing observations and imbalanced classes) approaches for future air quality forecasting. The publication is well structured, theoretical calculations have been tested on air quality prediction based on data set of the India AQI.

Submission Date
28 February 2023
Date of this review
22 Mar 2023 13:26:17