Testing Three Techniques of Handling Missing Values

Parth D Thummar
Department of Information & Technology Engineering
A. D. Patel Institute of Technology New Vidya Nagar, Anand, Gujarat, India

Abstract— nowadays, a plenty of data is generated each and every day and to retrieve useful and desired information is getting crucial for data scientist. Due to some reasons incomplete data sets have been seen while data mining and it is very important to handle that missing values in data set so that the data can be mined effectively. So here i am going to show three methods of handling missing values and compare the accuracy of each one on small data set.

Keywords: Datamining, Mising values, Handling techniques, Accuracy, Effectiveness, Naïve bayes algorithm, ingnore tuple, central tendency

I. Introduction

Missing values are common in today's era. Although it looks small, it has very large impact on data mining sometimes. Data set is incomplete in the sense of lacking attribute values or certain attributes of interest, it may contain only aggregate data. There are several reasons by which dataset remains incomplete. One of the popular reason is by human typing error; sometimes typing in hurry may cause incomplete data or wrong entry of data. Sometimes equipment error also cause to incomplete data. The motive behind this paper is to handle missing values using three techniques and test the accuracy of each and every methods on small data sets.

II. Method Overview

First and foremost, each method has it's own way to fill the data with specific values. Now we already have three small data sets and then we handles missing values in them and finally we will decide the accuracy of them by applying some query on it.

A. Ignore the whole row or column:

In this method the row or column of the dataset which has missing values is ignored or deleted. It may possible that dataset contains so many rows or columns with missing values; then all the affected are to be deleted. This usually done when class label is missing.

Example: Consider the following data.

Black	new	Damaged
Blue	old	?

In the above case the whole row containing missing value is deleted and then our data looks like:

Black	new	Damaged

B. Use central tendency:

In this method missing values are fulfilled by values which has central tendency in the dataset like most repeating values, mean, median Etc. Sometimes filling the values according to the same class also possible and proven to be more efficient.

Example:

Black	new	Damaged
Blue	old	Not damaged
Black	new	?

In this case the missing value has class black and we have seen that in the first row that black car is damaged so we have to put Damaged in the place of missing values.

Black	new	Damaged
Blue	old	Not damaged
Black	new	Damaged

C. Use most probable value with naïve Bayesian formula:

In this method all the missing values are handled by applying the algorithm and putting the most probable calculated value. This method is working on the bases of conditional independence.

According to bayesian theorem,

$$P(H|X)=P(X|H)P(H) \ / \ P(X)$$

Example:

Outlook	Temperature	Humidity	Play cricket
Rainy	Hot	High	NO
Rainy	cold	Low	Yes
Rainy	Hot	Low	Yes
Rainy	cold	High	No
Rainy	cold	High	?

Probability of yes
=P(Rainy|Yes)*P(cold|yes)*P(High|yes)*P(yes)
=2/4 * 1/2 * 0/2 * 2/4 =0
Probabilty of No =
P(Rainy|No)*P(cold|No)*P(High|No)*P(No)
=2/4 * 1/2* 1/1 * 2/4 =0.125

Probability of No is higher than probability of yes hence we have to put NO in the place of YES; our data looks like:

Outlook	Temperature	Humidity	Play cricket
Rainy	Hot	High	No
Rainy	cold	Low	Yes
Rainy	Hot	Low	Yes
Rainy	cold	High	No
Rainy	cold	High	No

III. Proposed Work

A. Dataset 1:

1) Original dataset:

ID	COLOR	WEIGHT	BROKEN
1	Black	250	Yes
2	Yellow	250	No
3	Blue	300	Yes
4	Black	250	Yes
5	Yellow	200	No
6	Black	250	Yes

2) Now suppose the missing values are following

ID	COLOR	WEIGHT	BROKEN
1	Black	250	Yes
2	Yellow	250	No
3	?	300	Yes
4	Black	250	Yes

5	Yellow	200	?
6	Black	250	Yes

3) After applying ignore the tuple method dataset looks like:

ID	COLOR	WEIGHT	BROKEN
1	Black	250	Yes
2	Yellow	250	No
4	Black	250	Yes
6	Black	250	Yes

4) Based on this new database selected queries and their answers are as following:

1) Which color watch has highest weight?
   Answer: Black, Yellow
2) Which color watch tend to breakable more?
   Answer: Black
3) Which color watch has lowest weight?
   Answer: Black, Yellow
4) Which color watch tend to breakable less?
   Answer: Yellow
5) Which color watch purchased less?
   Answer: Yellow

5) After applying Central tendency method our dataset looks like:

ID	COLOR	WEIGHT	BROKEN
1	Black	250	Yes
2	Yellow	250	No
3	Black	300	Yes
4	Black	250	Yes
5	Yellow	200	No
6	Black	250	Yes

Answers of the above queries are updated and as following:
1. Black 2.Black 3.Yellow 4.Yellow 5.Yellow

6) After applying naïve bayesian algorithm our dataset looks like:

After applying this method we are getting the result same as the above method, Hence all the results are remain same as previous method.

B. Dataset 2:

1) Original dataset looks like:

Age	Income	Student	Buy_computer
Youth	High	No	No
Youth	High	Yes	No
Middle_age	High	No	Yes
Senior	Medium	No	Yes
Senior	Low	Yes	No
Middle_age	Low	Yes	Yes
Youth	Medium	No	No
Senior	Medium	Yes	Yes
Middle_age	High	Yes	Yes
Middle_age	Medium	No	Yes

2) Now suppose the missing values are following:

Age	Income	Student	Buy_computer
Youth	High	No	No
Youth	High	Yes	No
Middle_age	High	?	Yes
Senior	Medium	No	Yes
Senior	Low	Yes	?
Middle_age	Low	Yes	Yes
Youth	?	No	No
Senior	Medium	Yes	Yes
Middle_age	High	Yes	Yes
Middle_age	Medium	No	?

3) After applying ignore the row method our dataset looks like:

Age	Income	Student	Buy_computer
Youth	High	No	No
Youth	High	Yes	No
Senior	Medium	No	Yes
Middle_age	Low	Yes	Yes
Senior	Medium	Yes	Yes
Middle_age	High	Yes	Yes

4) Based on this new database selected queries and their answers are as following:

1) Chances of buy_computer= yes when income=high and student= no?
   Answer: 0%
2) Chances of student= yes when age=youth and income= high?
   Answer: 50%
3) Chances of buy_computer= no when income=low and student= yes?
   Answer:0%
4) Chances of income= medium when age=senior and student= yes?
   Answer: 100%
5) Chances of buy_computer= yes when income=medium and student= no?
   Answer: 100%

5) After applying central tendency method our dataset looks like:

Age	Income	Student	Buy_computer
Youth	High	No	No
Youth	High	Yes	No
Middle_age	High	Yes	Yes
Senior	Medium	No	Yes
Senior	Low	Yes	Yes
Middle_age	Low	Yes	Yes
Youth	high	No	No
Senior	Medium	Yes	Yes
Middle_age	High	Yes	Yes
Middle_age	Medium	No	Yes

Answers of the above queries are updated and as following:
1)0% 2)33.33% 3)0% 4)50% 5)100%

6) After applying naïve bayesian algorithm our dataset looks like:

After applying this method we are getting the result same as the above method, Hence all the results are remain same as previous method.

C. Dataset 3:

1) Original dataset looks like:

Chills	Runny nose	Headache	Fever	Flue
Yes	no	Mild	Yes	No
Yes	Yes	No	No	Yes
Yes	No	Strong	Yes	Yes
No	Yes	Mild	Yes	Yes
No	No	No	No	No
No	Yes	Strong	Yes	Yes
No	Yes	Strong	No	No
yes	Yes	Mild	Yes	Yes

2) Now suppose the missing values are following:

Chills	Runny nose	Headache	Fever	Flue
Yes	no	Mild	Yes	No
Yes	Yes	No	No	Yes
Yes	No	Strong	Yes	?
No	?	Mild	Yes	Yes
No	No	No	No	No
No	Yes	Strong	Yes	Yes
No	Yes	Strong	No	?
yes	Yes	?	Yes	Yes

3) After applying ignore the row method our dataset looks like:

Chills	Runny nose	Headache	Fever	Flue
Yes	no	Mild	Yes	No
Yes	Yes	No	No	Yes
No	No	No	No	No
No	Yes	Strong	Yes	Yes

Based on this new database selected queries and their answers are as following:

1) Chances of flue= yes when fever=yes, headache= strong, chills= yes?
   Answer: 0%
2) Chances of headache= no when fever=no, flue= yes, runny nose=yes?
   Answer: 100%
3) Chances of flue= yes when runny nose= yes, headache= strong, fever=no?
   Answer:0%
4) Chances of runny nose= yes when headache=mild, fever= yes?
   Answer: 0%
5) Chances of fever= yes when flue= yes, headache= strong?
   Answer: 100%

4) After applying central tendency method our dataset looks like:

Chills	Runny nose	Headache	Fever	Flue
Yes	no	Mild	Yes	No
Yes	Yes	No	No	Yes
Yes	No	Strong	Yes	Yes
No	Yes	Mild	Yes	Yes
No	No	No	No	No
No	Yes	Strong	Yes	Yes
No	Yes	Strong	No	Yes
yes	Yes	Strong	Yes	Yes

Answers of the above queries are updated and as following:
1)100% 2)50% 3)100% 4)50% 5)75%

5) After applying naïve bayesian method our dataset looks like:

Chills	Runny nose	Headache	Fever	Flue
Yes	no	Mild	Yes	No
Yes	Yes	No	No	Yes
Yes	No	Strong	Yes	No
No	Yes	Mild	Yes	Yes
No	No	No	No	No
No	Yes	Strong	Yes	Yes
No	Yes	Strong	No	Yes
yes	Yes	Strong	Yes	Yes

Answers of the above queries are updated and as following:
1)50% 2)50% 3)100% 4)50% 5)66.66%

D. Efficiency:

Now efficiency of the three methods are checked based on how many queries answers are correct with respect to original dataset. Following table indicates the accuracy in percentage.

1) For database 1:

	Q1	Q2	Q3	Q4	Q5	TOTAL
Method 1	false	true	false	true	false	40%
Method 2	false	true	true	true	false	60%
Method 3	false	true	true	true	false	60%

2) For database 2:

	Q1	Q2	Q3	Q4	Q5	TOTAL
Method 1	false	true	false	false	false	20%
Method 2	false	false	true	true	false	40%
Method 3	false	false	true	true	false	40%

3) For database 3:

	Q1	Q2	Q3	Q4	Q5	TOTAL
Method 1	false	true	true	false	true	60%
Method 2	true	false	false	false	false	20%
Method 3	false	false	false	false	false	0%

4) Average of all three database:

Method 1	40%
Method 2	40%
Method 3	33.33%

IV. CONCLUSION

To recapitulate, all three methods shows almost equal efficiency averagely. One cannot say that particular method is feasible; because each method has its own importance and it is fully dependent on what type of database is used. For first method ignore the row or column; we can say that if database has limited or small amount of missing values then and then only this method should be used otherwise it will delete a lot of data. For second method we can say that if database contain limited class attributed then this method is efficient and for third method, if missing class attributes values have only two values then it will efficient. In short uses of method completely relies on number of missing values and type of the database.

REFERENCES

[1] Jiří Kaiser proposed a paper on "Dealing with Missing Values in Data"
[2] Edgar Acun˜a and Caroline Rodriguez proposed a paper on, "The treatment of missing values and its eff ect in the classifier accuracy"
[3] Arnaud Ragel and Bruno Cr6milleux proposed a paper Treatment of Missing Values for Association Rules"
[4] Marvin L. Brown and john F. Kros proposed a paper, "Data mining and the impact of missing data"
[5] Jerzy W. Grzymala-Busse and Witold J. Grzymala-Busse proposed a paper, "handling missing attribute value"

