CENG 414
Introduction to Data Mining
Fall 2019-2020
THE 2

Question 1

1- Which method(s) did ReplaceMissingValues use to replace missing values?

ReplaceMissingValues filter uses mean or mode of the preprocessed attributes. There are 17 attributes in the “labor.arrf” file. By looking mean values of attributes both before and after the application of ReplaceMissingValues, mean of the attributes is not changed. If data is nominal, then ReplaceMissingValues assigns all missing values as most frequent value which is mode. Thus, all replaced values for missing ones are mean or mode values corresponding to it’s data type.

2- When you compare the statistics of all attributes after applying the function with the raw statistics, which statistic(s) have changed?

Minimum, maximum and mean values were not changed after applying the ReplaceMissingValues filter. However, number of missing values, number of distinct values and most importantly, standard deviation of the data are changed.

3- How did the dataset be affected after applying ReplaceMissingValues function in terms of the changes in its statistics ? Discuss briefly for the following attributes only: ”duration”,”standby-pay”,”wage-increase-third-year”,”wage-increase-first-year”.

There is 57 instances in the “labor.arrf”. If a lot of new values are added by ReplaceMissingValues function, than that changes attribute exceedingly. If missing value number is relatively less than total instances, this filter may be useful.

For the duration attribute:

Since there is only 1 missing value in duration attribute, replacing 1 missing value with attribute’s mean will not change the characteristics of data. Statistically,

- ➔ mean won’t be changed since replaced value is also mean. Before applying it was 2.161 and after applying ReplaceMissingValues it still remains 2.161.
- ➔ minimum and maximum won’t be changed since replaced value mean is between min and max values. Before applying ReplaceMissingValues,min

was 1 and max was 3.After applying ReplaceMissingValues, their values are still same.

- ➔ Standard deviation is changed a little bit. It's value is 0.708 before ReplaceMissingValues filter and after applying this filter to duration attribute, it becomes 0.701. There should be a reduction in standard deviation since more values exist around the mean after adding one value(which equals to mean) to attribute.

For the standby-pay attribute:

There are 48 missing values in standby-pay attribute and that means %84 of the values are not exist. Replacing 48/57 missing values will be able to change the characteristics of data totaly.

Statistically,

- ➔ mean won't be changed since replaced values are also mean. Before applying it was 7.444 and after applying ReplaceMissingValues it still remains 7.444.
- ➔ minimum and maximum won't be changed since mean is between min and max values. Before applying ReplaceMissingValues,min was 2 and max was 14. After applying ReplaceMissingValues, their values are still same.
- ➔ Standard deviation is changed a lot. It's value is 5.028 before ReplaceMissingValues filter and after applying this filter to standby-pay attribute, it becomes 1.9.

A lot new values are added and they changed the characteristics of the attribute exceedingly.

For the wage-increase-third-year attribute:

There are 42 missing values in wage-increase-third-year attribute and that means %74 of the values are not exist. Replacing 42/57 missing values will be able to change the characteristics of data totaly.

Statistically,

- ➔ mean won't be changed since replaced values are also mean. Before applying it was 3.913 and after applying ReplaceMissingValues it still remains 3.913.
- ➔ minimum and maximum won't be changed since mean is between min and max values. Before applying ReplaceMissingValues,min was 2 and max was 5.1. After applying ReplaceMissingValues, their values are still same.
- ➔ Standard deviation is changed a lot. It's value is 1.304 before ReplaceMissingValues filter and after applying this filter to wage-increase-third-year attribute, it becomes 0.701.

A lot new values are added and they changed the characteristics of the attribute exceedingly.

For the wage-increase-first-year attribute:

Since there is only 1 missing value in duration attribute, replacing 1 missing value with attribute's mean will not change the characteristics of data. Statistically,

- ➔ mean won't be changed since replaced value is also mean. Before applying it was 3.804 and after applying ReplaceMissingValues it still remains 3.804.
- ➔ minimum and maximum won't be changed since replaced value mean is between min and max values. Before appliying ReplaceMissingValues,min was 2 and max was 7. After applying ReplaceMissingValues, their values are still same.
- ➔ Standard deviation is changed a little bit. It's value is 1.371 before ReplaceMissingValues filter and after applying this filter to duration attribute, it becomes 1.358.
- ➔ There should be a reduction in standard deviation since more values exist around the mean after adding one value(which equals to mean) to attribute.

4- Is ReplaceMissingValues function suitable to replace missing values? Discuss briefly for the following attributes only: "duration","standby-pay","wage-increase-third-year","wage-increase-first-year". If you think it is not suitable for some attribute(s), briefly discuss why.

For the duration attribute:

There is only 1 missing value in duration attribute. Replacing 1 missing value with attribute's mean will not change the characteristics of data. However, all values in the duration attribute are integers. If there is not a constraint on this attribute like all attributes should be integer, ReplaceMissingValues function is suitable for duration attribute.

For the standby-pay attribute:

There are 48 missing values in standby-pay attribute and that means %84 of the values are not exist. Replacing 48/57 missing values will be able to change the characteristics of data exceedingly. Thus, this filter is not suitable for standby-pay attribute.

For the wage-increase-third-year attribute:

There are 42 missing values in wage-increase-third-year attribute and that means %74 of the values are not exist. Replacing 42/57 missing values will be able to change the characteristics of data exceedingly. ReplaceMissingValues function is not suitable for this attribute as well.

For the wage-increase-first-year attribute:

Since there is only 1 missing value in duration attribute, replacing 1 missing value with attribute's mean will not change the characteristics of data totaly. But like in the duration attribute, if there is a constraint on values that they should be integer, this method is not suitable since replaced value is not an integer(it is mean and it is 3.804). Otherwise, this method is suitable.

Alper KOCAMAN - 2169589