Cluster Based Knn Imputation

Eviatar Nachshoni[1], Yosef Danan[1]

[1]Bar Ilan University
{eviatarn, yosefdanan1234}@gmail.com

February 2022

Abstract

Filling null values in data observation is one of the major steps in data science pipeline. This is even more crucial for cases where there have small datasets. In our research we focus on the KNN imputation method, but instead to determine for each sample the same k, we use clustering method to adapt for each of the samples with nan value the cluster which he most probably from there and run the KNN on this and only this neighbours. we run our method on multiple datasets and we found that our method improve the results in cases where the percentage of null values is high.

1 Problem description

The problem of missing (or ”NaN” for ”not a number”) values occurs when a dataset contains empty or null values. This can be problematic for a number of reasons:

1. Filling missing values can introduce bias into the analysis, particularly if the missing data is not missing at random.

2. Filling missing values can also inflate the variance of the data and lead to incorrect standard error estimates, which can have a negative impact on the inferences and conclusions drawn from the data.

All these problems lead to try to lower the MSE for the null values over all the samples that contain null values.

2 Related Work

In the area of null imputation have a lot of works, where for each of the method have her pros and cons, here is part of them:

1. Mean or median[1]:
   Replacing missing values with the mean or median of the non-missing values in the same column.
   They have some limitation[1]:

   (a) They assume a normal distribution of the data which might not be the case in the major of scenarios.

   (b) It's also important to note that these methods can lead to biased and inefficient estimations of parameters.

2. MICE[3]:
   Creating multiple "imputed" datasets by using a model to generate plausible values for missing data, and then analyzing each dataset separately and combining the results. They disadvantage of this method:

[1] Each of the methods has multiple disadvantages, but we will relate to a few of them.

   (a) MICE might not perform well when the amount of missing data is large.

3. K-nearest neighbors[4]:

   The similarity between observations is determined by their distance in the feature space, and the k-nearest neighbors are the observations that have the smallest distance to the observation with missing values. There have two different variant of this method:

   (a) Uniform method: All points in each neighborhood are weighted equally.

   (b) Distance method: weight points by the inverse of their distance. in this case, closer neighbors of a query point will have a greater influence than neighbors which are further away [5].

   One of the limitations of the KNN imputation method is that the same k value is often used for all samples, regardless of the density of observations in the feature space. This can lead to some problems:

   (a) The k value might be too small in areas of high density, resulting in imputed values that are not representative of the true values.

   (b) The k value might be too large in areas of low density, resulting in imputed values that are not representative of the true values.

3 Solution overview

The k-nearest neighbors (KNN) imputation method can be improved by incorporating clustering techniques. Clustering is a method of grouping similar observations together based on their characteristics. By grouping similar observations together, clustering can help to identify regions of high density in the feature space, which can be used to improve the KNN imputation method.

In order to estimate the missing value, instead of looking at a constant k that we received from the user, we will use the compression of each cluster, that is, the more compressed a cluster is, the greater the number of neighbors, and vice versa

There are two ways to evaluate each imputation method:

1. make nan of features that we already know his actual value and calaculate the mean squared error between the real value and the imputed value

2. mpute real missing value and run a baseline classification model on imputed values and see who gives the best results on the end-problem(accuracy, precision, recall etc.)

Algorithm 1 Cluster Based Knn Imputation

```
Require: let define $k_i \in K$ to be range of k which get from the user, and let define S to be range of number of clusters we check upon the clustering algorithms,
  $x_i \in X$ be a data-set where $x_i$ is single sample ,and let $y_i \in Y$ be samples which contain null values, where $Y \subset X$
Ensure: let define T to be $X \setminus Y$
  let $M = \{$K-means, GMM, Hirarchial Clustering, DBSCAN$\}$
  for each $m_i \in \mathcal{M}$ do
    for each $s_i \in S$ do
      run $m_i$ on T, and then get $c_{ij} \in C$ where $i$ define the clustering method and $j$ define the number of clusers under the specific method.
    end for
    keep the $s_i$ which $max(silhoutte(M_i))$
  end for
  Extract the method which maximise the silhoutte.
  for each $y_i \in \mathcal{Y}$ do
    find the $c_i \in C$ which $max(\forall x_i \in C_i \; cos(y_i, xi))$
    let $D =$ density of each $c_i \in C$
    let $knn_i = lower(\frac{(d_i - min(D)) \times (k - 0.1k)}{max(D) - min(D)} + 0.1k))$
    let $N = knn_i$ neighbores of $y_i \in c_i$
    $$y_i = \left\{ \begin{array}{cc} \text{mean}(N), & \text{Numerical feature} \\ \text{mode}(N), & \text{Categorical feature} \end{array} \right\}$$
  end for
```

4 Experiments

In order to gauge the quality of the method we use, we tested it on several different dataset:

Name	No.Columns	No. Samples	Explanation
Titanic	8	712	passengers details and survival status
Mobile	21	2000	mobile phone specifications and prices
Mnist	64	1797	handwritten digits gray scale images

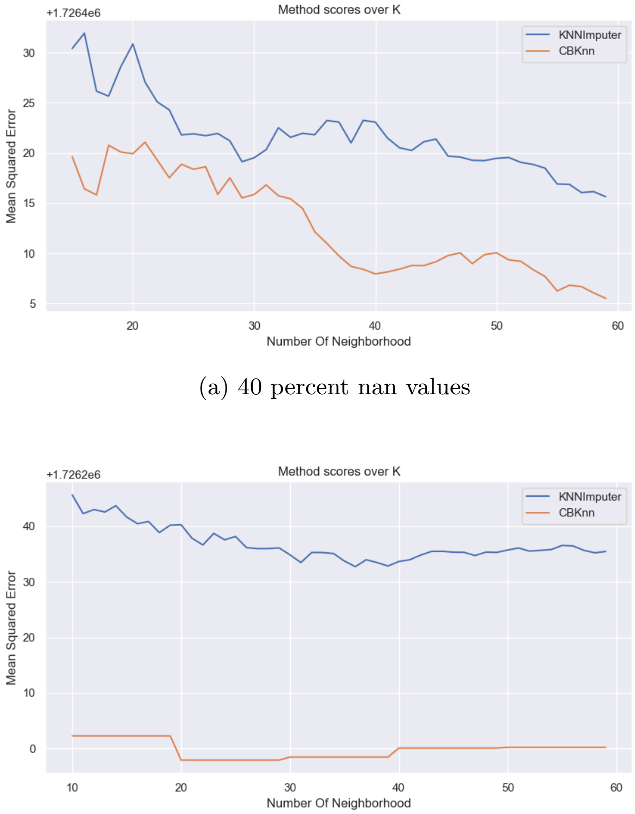
Table 1: Datasets details

For each of the datasets we checked several parameters:

1. Different amount of missing values:
   We tested different percentages of missing values, what we saw is that the more missing values the methods converge to be equal or at least to be similiar, and this is expected, because it is more difficult to complete from each cluster, since they become smaller.

2. We removed missing values using different distributions:
   At first we tested samples with equal probability, and then we tested samples with unequal probability, and we saw that for unequal probability, the algorithm may be less good, as expected again, since we are looking inside the same cluster and now there are more missing values, so it is difficult to fill in the missing values in that cluster well.

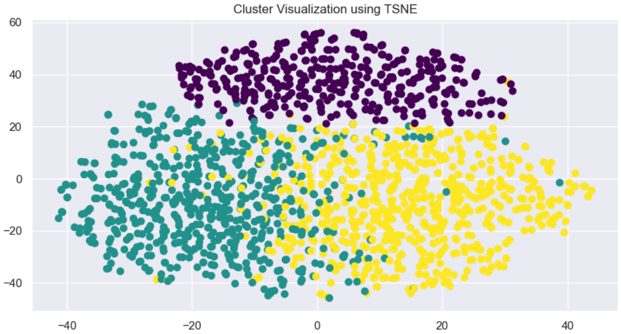
1. Results graphs on mobile dataset:



(a) 40 percent nan values

(b) 20 percent nan values

Figure 1: Mobile datsets results



(a) TSNE

Figure 2: Clustring Visualization

5 Conclusion

In the case of quality clustering (measured by silhouette), we could use clustering methods to fill in missing values more efficiently. Using clustering methods, we determined the compression of each cluster and determined the K for each missing value based on the compression. The methods we used to fill the missing values improved the quality of the filling missing values.

References

[1] A.Rogier, T.Donders, Geert J.M.G., van der Heijden, Theo Stijnen Karel G.M.Moonsc (2006), A gentle introduction to imputation of missing values

[2] Little, R.J., Rubin, D.B. (1987) ”Random sample imputation for missing data: A comparison with mean and regression imputation”

[3] van Buuren, S., Groothuis-Oudshoorn, K. (2011). ”Multiple imputation for nonresponse in surveys”

[4] El-Shazly, A., Al-Zubaidy, H. (2019). ”A Comparative Study of KNN and Mean Imputation for Missing Values in Medical Datasets”.

[5] Stef van Buuren and Karin Groothuis-Oudshoorn (2011) ”A Comparative Study of K-Nearest Neighbors Imputation and Multivariate Imputation by Chained Equations”.