Final-Project - CBKnn Imputation

Yosef Danan, Eviatar Nachshoni

March 2023

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 project we focus on the KNN imputation method, but instead to determining the same k which we got from the user for every sample, we use a clustering method to adjust for each sample with a nan value the cluster from which he is likely to originate and evaluate the clustering density on the cluster. Thus, as the density increases, the k will increase. we run our method on multiple datasets and compare it with classic KNN imputation, we found that our method improve the results in cases where the percentage of null values is high and there have high quality of clustering.

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:

   Replacing missing values with the mean or median of the non-missing values in the same column. The problem with this method is[1] that they assume a normal distribution of the data which might not be the case in the major of scenarios.

2. MICE[2]:

   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:

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

3. K-nearest neighbors[3]:

   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 [4].

   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.

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

(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 change the K depending on the compression of each cluster, that is, the more compressed a cluster is, the greater the number of neighbors, and vice versa

The intuition for this assumption is, when a group of samples is very dense, it is an indication that they are probably from the same distribution, so we would like to estimate the missing value by using all the samples we have, although, if the cluster is very sparse, this means that the samples do not necessarily come from the same distribution, therefore We will reduce the size of the k that came from the user.

Algorithm 1: Cluster Based Knn Imputation

```
Require: Let define $k$ to be number of neighborhood 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 $m_i$ and $s_i$ on T.
      end for
  end for
  keep the $m_i$ and $s_i$ which $max(silhoutte(m_i, s_i))$
  Let define $c_i \in \mathcal{C}$ to be the clusters we got from the $m_i$ and $s_i$.
  Let $d_i \in D$ to be the density of $c_i \in C$
  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 $k_i = lower(\frac{(d_i - min(D)) \times (k - 0.1k)}{max(D) - min(D)} + 0.1k))$
      Let $N =$ neighbores of $y_i \in c_i$ we got from KNN algorithm with $k_i$
          $y_i = \left\{ \begin{array}{ll} \text{mean}(N), & \text{Numerical feature} \\ \text{mode}(N), & \text{Categorical feature} \end{array} \right\}$
  end for
```

There are two ways to evaluate each imputation method:

1. Extract randomly feature which not contain NaN values, and manually create NaN values, and keep the real values, then, run the imputation method and calculate the mean squared error between the real value and the imputed value

2. Impute 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.)

In our project we focus on the first way, and we use the classic KNN imputation as the baseline model.

4 Experiments

In order to gauge the quality of the method we use, we tested our method on several different which apear in Table 1.

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
Iris	5	150	Properties for each of the flowers

Table 1: Datasets details

For each of the datasets we choose one of the columns and checked several parameters:

1. Different amount of missing values:
   We tested different percentages of missing values [20%, 40%], and we keep this values ad the ground truth.

2. We removed missing values using different distributions:
   At first we removed values from the column with equal probability, and then we tested samples with unequal probability.

   Then we run our datasets with the different cases using multiple step and evaluate our model using the KNN imputation method as baseline. First, we choose the best clustering algorithm using silhouette method over different size of clustering we already insert. Then, for the best clustering algorithm, we rechange the K which we got from the user based on the density of the cluster [Algorithm 1]. We found that the quality of the clustering greatly affects the quality of our result, i.e. in datasets where it is difficult to build quality clusters, such as in Iris and Mnist, our results also worked less well, but when we got high quality clustering for Titanic as show from Figure 1, we can see that the results in compare to KNN imputation are beteer as shown from Figure 3. we too see that for Titanic dataset, where we succeed to get high quality clustering, we succeed to get better results in compare to the classic KNN imputation method.

We too inspect that for different clustering algorithms the results compared to the classic KNN were sometimes good and sometimes less, which support our conception that the quality of the clustering, highly affect on our method.

In addition, we saw that the more the amount of data that is missing, the better our method is compared to a regular KNN as shown in Figure 3_b in compare to 3_a . Also, sampling with a very different distribution affects the model, when sampling with a uniform distribution, so when the amount of nan is large, our method works better. we think that in non normal distribution can hurt the clustering algorithm quality which highly affect on our method.

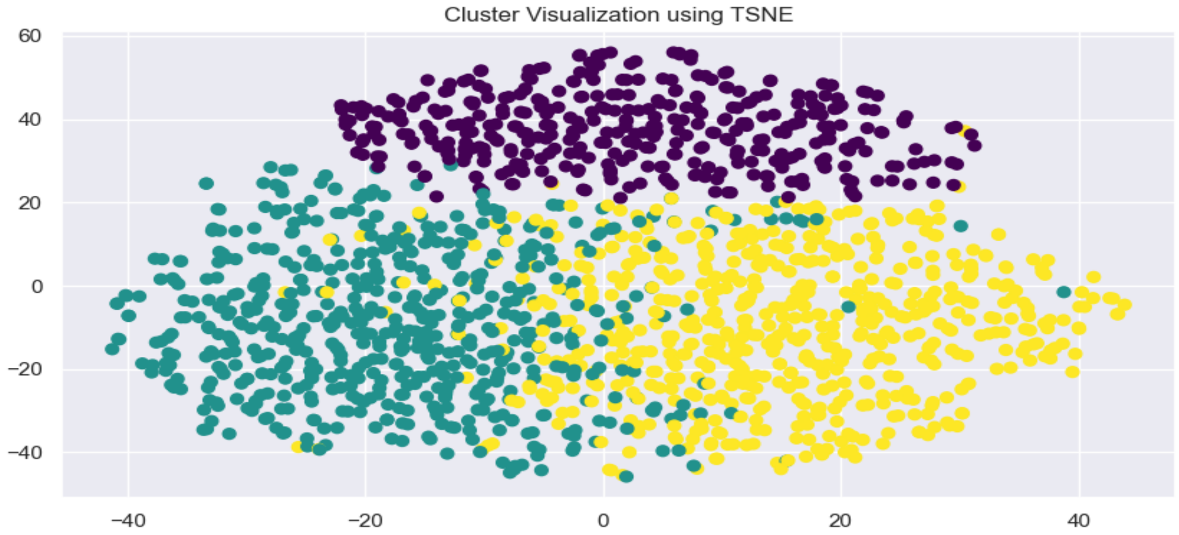
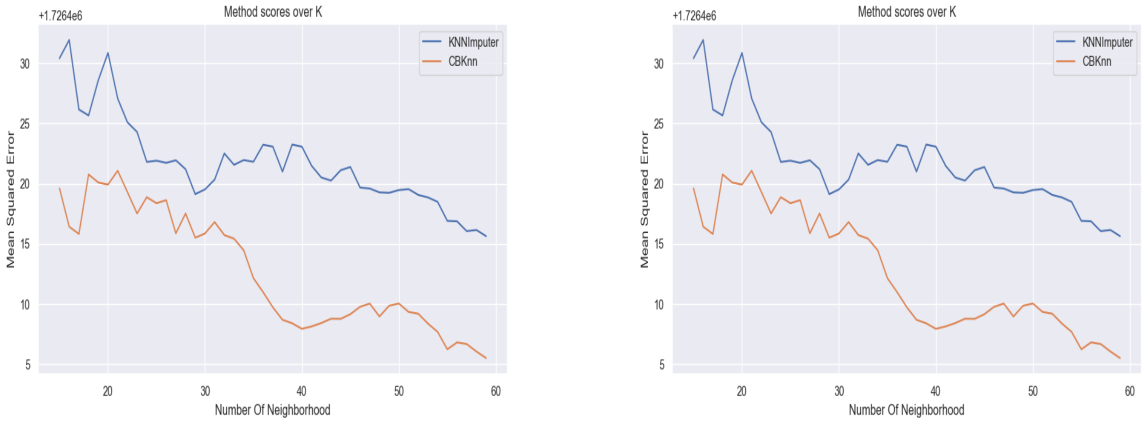


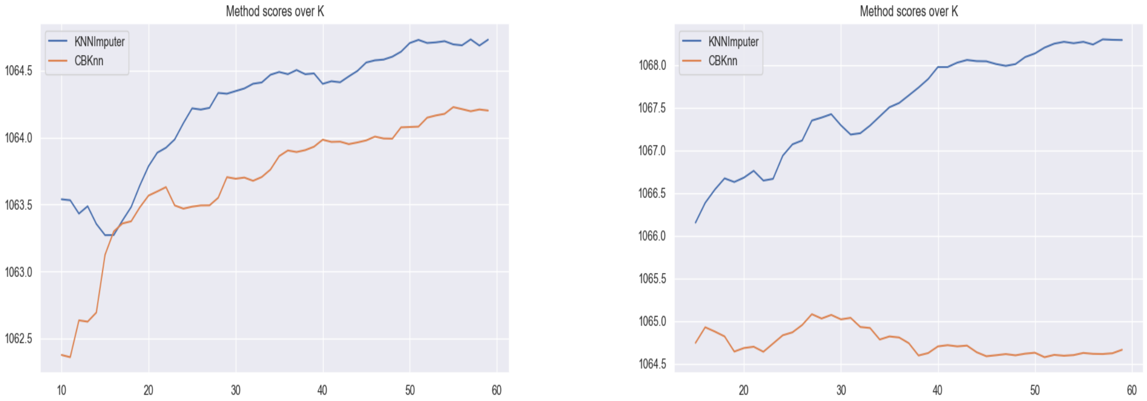
Figure 1: Clustring Visualization Mobile dataset



(a) 20 percent nan values

(b) 40 percent nan values

Figure 2: Titanic dataset Evaluation



(a) 20 percent nan values

(b) 40 percent nan values

Figure 3: Mobile dataset Evaluation

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, which enable us to vary K in compare to the classic KNN imputation method. The methods we used to fill the missing values improved the quality of the filling missing values when the quality of the clustering was good.

References

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

[2] https://www.numpyninja.com/post/mice-algorithm-to-impute-missing-values-in-a-dataset

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

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