MACHINE LEARNING FOR MISSING-DATA IMPUTATION

E. Ben-Akoune, B. Lerner and A. Even
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

ABSTRACT
Missing values are a common, and often hazardous, type of data quality defect in organizational data repositories. Recently, machine learning and data mining algorithms addressed missing-value issues by the automation of the data-imputation process. We use the $k$-nearest neighbor ($k$NN) algorithm for missing-data imputation, and evaluate two criteria for measuring the imputation success – imputation distance, measuring the similarity between the imputed and the real-world data, and classification accuracy due to imputation. Using several test datasets, we show that usually the imputation distance increases with the missing-data rate. Also, the $k$NN classification accuracy is almost always higher for the imputed data than for the missing data, whereas the naïve Bayesian classifier shows opposite results. We suggest explanation to this inconsistency.

1. INTRODUCTION
Missing values are among the most common and hazardous defects in organizational data repositories. They can be the result of different causes, such as the unavailability or irrelevance of certain pieces of data when a data record is first recorded, a failure to include required values by the person who enters or audits the data, a mismatch with the associated value domain which prevents data from being stored in the database, or a technical failure which causes the data value to be deleted by mistake.

Since decisions that are based on data with significant numbers of missing values might turn out to be flawed, organizations invest vast amounts of time, money and human resources in correcting such defects. This is often done manually, by auditing the dataset records and filling-in the missing values. As manual imputation is timely and costly, studies have explored methods for automated data imputation [1]. The goal of this study is to develop and explore an automated data-imputation method based on the $k$-nearest neighbor ($k$NN) algorithm to improve the quality of organizational data.

2. THEORETICAL BACKGROUND
A test data instance is classified using $k$NN according to the most common class label among the instance $k$-nearest training neighbors. The nearness between instances is defined in
terms of distance, which increases with the dissimilarity between the instances. For choosing
the nearest neighbors correctly, the distance metric must consider linear (i.e., continuous or
discrete) and nominal data attributes having different unit scales. Such is the Heterogeneous
Value Difference Metric (HVDM) [2]:

\[
HVDM(x_m, x_l) = \sqrt{\sum_{r=1}^{2} d(a_r(x_m), a_r(x_l))^2},
\]

where

\[
d(a_r(x_m), a_r(x_l)) = \begin{cases} 
1 & \text{if } a_r(x_m) \text{ or } a_r(x_l) \text{ are missing} \\
\sum_{c=1}^{C} \left( P(c_r | a_r(x_m)) - P(c_r | a_r(x_l)) \right)^2 \right)^{1/2} & \text{if } a_r \text{ is nominal} \\
\frac{|a_r(x_m) - a_r(x_l)|}{4\sigma_{a_r}} & \text{if } a_r \text{ is linear}
\end{cases}
\]

and \(a_r(\cdot)\) is the value of the \(r^{th}\) attribute (out of \(n\) attributes) of either a test instance \(x_m\) or a
training instance \(x_l\) for which HVDM is measured, \(P(c_r | a_r(x_l))\) is the posterior probability
of class \(c_r\) (out of \(C\) classes) given \(a_r(x_l)\), and \(\sigma_{a_r}\) is \(a_r\) standard deviation.

In this research, we develop a novel weighted \(k\)NN algorithm in which a weight is
assigned to the distance between test instance \(x_m\) and each training instance \(x_l\):

\[
w_l = \frac{1}{HVDM(x_m, x_l)^2}.
\]

Further, we use the weighted \(k\)NN algorithm to impute an attribute missing value in an
instance based on existing values in the instance \(k\)-nearest neighbors. The former instance is
acting like a “test instance” in the original \(k\)NN algorithm and the latter are the “training
instances”. Imputation is based only on existing values of the attribute in the \(k\)-nearest
neighbors. If some of the values are missing, the algorithm uses values from more distant
neighbors, such that the closest \(k\)-observed values are chosen. Imputation of the \(r^{th}\) attribute of
\(x_m\) by our novel method is performed by:
\[ a_r(x_m) = \begin{cases} \frac{\sum_{l=1}^{k} w_l \cdot a_r(x_l)}{\sum_{l=1}^{k} w_l} & \text{if } a_r(x_m) \text{ is continuous} \\ \frac{\left[ \sum_{l=1}^{k} w_l \cdot a_r(x_l) \right]}{\sum_{l=1}^{k} w_l} & \text{if } a_r(x_m) \text{ is discrete} \\ \arg\max_{a_r \in \text{Val}(a_r)} \sum_{l=1}^{k} w_l \cdot I(a_r(x_l), a_r) & \text{if } a_r(x_m) \text{ is nominal} \end{cases} \]

where \( I = 1 \) if \( a_r(x_i) = a_{r_j} \) and \( I = 0 \) if \( a_r(x_i) \neq a_{r_j} \). Our preliminary evaluation has shown that the suggested method has an advantage with respect to imputation.

3. EXPERIMENTAL METHOD

In our evaluation procedure, we first artificially implanted missing values at rates varying from 10 to 90 percents in three UCI datasets (Table 1) having no missing values, and then imputed the datasets by filling-in missing values using our novel method. The success of imputation was evaluated using two criteria:

1. Imputation distance that measures the similarity (distance) between the imputed dataset and the original dataset using HVDM. For perfect imputation, the imputation distance equals 0, and is greater otherwise.

2. Classification accuracy that measures the rate of accurate classification on a dataset using a 10-fold cross-validation experiment. The values of this criterion measured by a 10NN classifier (not to be confused with the kNN used for imputation) and the naïve Bayesian classifier (NBC) on the imputed dataset were compared to those measured on the missing dataset. That is, this measure demonstrates the contribution of imputation to classification accuracy of missing datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Instances</th>
<th>Nominal attributes</th>
<th>Linear attributes</th>
<th>classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>150</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Risk</td>
<td>1000</td>
<td>6</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Chess</td>
<td>1000</td>
<td>36</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
4. RESULTS

Table 2 shows that the $k$NN classification accuracies on the imputed datasets are higher than on the missing datasets. For the Risk and Chess datasets, this result was consistent at any missing-value rate, and for the Iris dataset at any rate, but 10 and 20 percents. On the contrary, the NBC classification accuracy was consistently better on the missing datasets than on the imputed datasets. This result complies with a study by Liu et al. [3], which found NBC to be robust with missing data. This NBC robustness – mainly due to the effect of likelihood decomposition on the curse of dimensionality and on correct classification – can make imputation unnecessary and even harmful when using NBC.

A comparison of the $k$NN accuracy on the imputed data with the NBC accuracy on the missing data for Risk and Chess highlights an interesting pattern. For missing rates up to around 30 percent, the former is higher due to the strength of imputation. However, for rates higher than around 30 percent, the latter dominates, as the NBC accuracy does not deteriorate substantially as that of $k$NN. For all three datasets, the classification accuracy, measured on the missing or imputed datasets, decreases, as expected, with the missing rate.

Figure 1 shows the imputation distance for increasing missing rates for the datasets. While the distance rises moderately for Risk and steeply for Iris, it seems almost independent on the rate for Chess. The lack of dependence for Chess can be explained by the use of HVDM as also discussed in [3]. HVDM for a nominal attribute measures its ability to discriminate between classes. For a non-discriminative attribute (i.e., an attribute that do not differentiate well between classes), the applied range of values will be relatively small, and thus missing values are less likely to contribute to the imputation distance substantially. It can be shown that the Chess dataset, which consists of 36 nominal attributes, has only a few discriminative attributes. This can possibly explain why the effect of missing rate on the imputation distance in this dataset is relatively small. Nevertheless, the increase in the missing rate still affects the discriminative attributes and thereby undermines classification accuracy (Table 2).

5. CONCLUSIONS

This study examined the accuracy of the $k$NN algorithm as an imputation method. For the tested datasets, we have found that for missing rates of up to around 30 percent, classification accuracy using $k$NN on imputed data outperforms that using NBC. However, for higher missing-value rates, NBC achieves better classification results, due to its robustness.

We currently investigate changing HVDM, such that it will treat missing values according to their importance to the classification task rather than equally. Another aspect is to consider
the correlations between attributes for imputing missing values more accurately. Currently, we use such correlations only between the class variable and the attributes.

### Table 2 Classification accuracy results (%)

<table>
<thead>
<tr>
<th>Classifier</th>
<th>% Missing</th>
<th>Iris</th>
<th>Risk</th>
<th>Chess</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Missing set</td>
<td>Imputed set</td>
<td>Missing set</td>
</tr>
<tr>
<td>kNN</td>
<td>0</td>
<td>95.2</td>
<td>95.2</td>
<td>74.9</td>
</tr>
<tr>
<td>NBC</td>
<td>10</td>
<td>87.6</td>
<td>87.6</td>
<td>71.3</td>
</tr>
<tr>
<td>kNN</td>
<td>20</td>
<td>90.4</td>
<td>92.6</td>
<td>74.0</td>
</tr>
<tr>
<td>NBC</td>
<td>30</td>
<td>87.6</td>
<td>87.5</td>
<td>70.8</td>
</tr>
<tr>
<td>kNN</td>
<td>40</td>
<td>90.4</td>
<td>88.4</td>
<td>71.8</td>
</tr>
<tr>
<td>NBC</td>
<td>50</td>
<td>88.1</td>
<td>88.7</td>
<td>70.7</td>
</tr>
<tr>
<td>kNN</td>
<td>60</td>
<td>88.9</td>
<td>90.4</td>
<td>66.6</td>
</tr>
<tr>
<td>NBC</td>
<td>70</td>
<td>85.5</td>
<td>83.7</td>
<td>71.0</td>
</tr>
<tr>
<td>kNN</td>
<td>80</td>
<td>85.0</td>
<td>86.0</td>
<td>64.0</td>
</tr>
<tr>
<td>NBC</td>
<td>90</td>
<td>80.7</td>
<td>76.1</td>
<td>69.0</td>
</tr>
<tr>
<td>kNN</td>
<td></td>
<td>81.6</td>
<td>82.0</td>
<td>60.5</td>
</tr>
<tr>
<td>NBC</td>
<td></td>
<td>78.3</td>
<td>76.7</td>
<td>67.9</td>
</tr>
<tr>
<td>kNN</td>
<td></td>
<td>73.3</td>
<td>78.4</td>
<td>58.5</td>
</tr>
<tr>
<td>NBC</td>
<td></td>
<td>71.7</td>
<td>68.7</td>
<td>65.8</td>
</tr>
<tr>
<td>kNN</td>
<td>40</td>
<td>61.4</td>
<td>63.8</td>
<td>56.0</td>
</tr>
<tr>
<td>NBC</td>
<td>50</td>
<td>60.6</td>
<td>57.7</td>
<td>64.4</td>
</tr>
</tbody>
</table>

![Figure 1 Imputation distance for increasing missing rates for Risk and Iris (left) and Chess (right)](image)

### REFERENCES


**ACKNOWLEDGEMENT:** The partial support for this research by the Ira Center for Business, Technology & Society, Ben-Gurion University, is gratefully acknowledged.