Archana Purwar* and Sandeep Kumar Singh

DBSCANI: Noise-Resistant Method for Missing Value Imputation

DOI 10.1515/jisys-2014-0172
Received November 22, 2014; previously published online July 10, 2015.

Abstract: The quality of data is an important task in the data mining. The validity of mining algorithms is reduced if data is not of good quality. The quality of data can be assessed in terms of missing values (MV) as well as noise present in the data set. Various imputation techniques have been studied in MV study, but little attention has been given on noise in earlier work. Moreover, to the best of knowledge, no one has used density-based spatial clustering of applications with noise (DBSCAN) clustering for MV imputation. This paper proposes a novel technique density-based imputation (DBSCANI) built on density-based clustering to deal with incomplete values in the presence of noise. Density-based clustering algorithm proposed by Kriegal groups the objects according to their density in spatial data bases. The high-density regions are known as clusters, and the low-density regions refer to the noise objects in the data set. A lot of experiments have been performed on the Iris data set from life science domain and Jain's (2D) data set from shape data sets. The performance of the proposed method is evaluated using root mean square error (RMSE) as well as it is compared with existing K-means imputation (KMI). Results show that our method is more noise resistant than KMI on data sets used under study.

Keywords: Missing value imputation (MVI), DBSCAN, K-means imputation, missing values, noise.

1 Introduction

Real-world databases are highly susceptible to noisy, missing, and inconsistent data. Missing values are present in many instances of data set if there is no recorded value for several attributes and noise is a random error or variances in a measured attribute or variable [11, 20]. These are attributed to their huge size and their origin from multiple heterogeneous sources [11]. Low-quality data leads to low-quality mining results. Missing values occur due to a number of reasons such as errors in the manual data entry procedures, equipment errors, or incorrect calculations. The presence of missing values (MVs) in data mining produces several problems in the knowledge extraction process such as loss of efficiency, complications in the management, and analysis of data. Data mining algorithms cannot be directly applied on the incomplete data set. Three types of techniques have been discussed in the literature to deal with the missing value. The first approach deletes the instances that are having missing entries. But this method fails when the data set consists of a huge number of incomplete values. The second kind of approach such as model based defines a model for the partially missing data and base inferences on the likelihood under that model. These models estimate the parameters by maximum likelihood, expected maximization, and other methods [13]. The third type of imputation techniques replaces the missing values by imputation such as the mean for numerical attributes and the mode for nominal attributes. These imputation techniques are independent of data mining algorithms used for the purpose of classification, clustering, or association. Hence, data miners can select any of the imputation techniques that suit their data set and apply learning algorithms on the imputed data set.

***Corresponding author: Archana Purwar,** Department of Computer Science and Information Technology, JIIT Noida, India, e-mail: archana.purwar@gmail.com
Sandeep Kumar Singh: Department of Computer Science and Information Technology, JIIT Noida, India

Although, a lot of work has been done in the area of missing value imputation, a slight attention except [1, 22, 25] has been given on noise in the present MV literature. The noise may come from many sources such as data transformation, data entry, or data compilation [23]. The presence of noise in the data can lead to false results during imputation. Therefore, there is a need to analyze the current imputation techniques in the presence of noise.

In this paper, we have proposed a novel imputation technique using density-based clustering approach and hypothesize that we can predict missing values correctly in noisy environment. Very few researchers have imputed missing values in the presence of noise. Density-based clustering algorithm [4] proposed by Kriegal groups the objects according to their density in spatial data bases. The high-density regions are known as clusters, and the low-density regions refer to the noise objects in the data set. A lot of experiments have been performed on two data sets, i.e. the Iris data set from the life science domain and 2D data set from shape data sets [2, 10]. We also have compared our results with the K-means imputation technique [13].

The rest of the paper is divided into four sections. Section 2 discusses the related work in MV literature. Section 3 presents the proposed work for imputation. Section 4 gives the experimental results and analysis drawn from them. Finally, Section 5 concludes the complete paper.

2 Related Work

The missing values present in the data set generate several problems like loss of efficiency and biased estimates in knowledge extraction processes. Different approaches for missing value imputation in MV literature are employed to prepare data in order to avoid the negative effects in the analysis of data mining algorithms. Broadly, they can be classified into the following categories:

- *List wise deletion:* The first and the most trivial approach is deleting the objects, which consists of incomplete entries. This method is based on list wise deletion of those examples, which have incomplete column values. But this approach is appropriate only when there are a couple of missing values in the data set. Although as the quantity of missing values increase, significant data may be lost by deletion of the incomplete instances [13].
- *Mean/mode substitution:* This is a simple way to impute the missing data. It replaces the missing values by the mean or mode of all the observations or a subgroup at the same variable. It consists of replacing the unknown value for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute. But replacing all missing records with a single value distorts the input data distribution [21].
- *Maximum likelihood procedures* [3]*:* These methods are used for estimation of parameters of model by making use of the expectation–maximization algorithm [13, 21]. These methods make assumption of model distribution for the variables, such as a multivariate normal model, which are very much sensitive to outliers [21].
- *Multiple imputation* [5, 18]: This method imputes the value of a missing variable more than once. Then, analysis is done by averaging them [21].
- *Machine learning-based imputation* [13]*:* There are a number of approaches for missing value imputation include choosing the most common value among the neighbors, weighted imputation based on k-nearest neighbors, decision tree based, association based, fuzzy logic, neural networks, SVM, and clustering-based imputation, e.g., as K-means, weighted K-means, fuzzy c means [6, 17, 19], etc. Recently, Ref. [24] has proposed an approach for mixed data set consisting of discrete as well as continuous attributes. Moreover, a novel clustering-based multiple imputations via Grey relational analysis has been stated in Ref. [21].

Most of the approaches discussed above ignore an important issue like noise that may be present in the data set in addition to missing values. The noise may arise from various resources such as data transformation, the process of data collection, data entry, etc. The presence of noise may introduce some harmful outcomes. A little work like robust imputation based on group method of data handling (RIBG) [1, 7–9, 16, 25] has been

done in the MV literature with noise. RIBG proposed by Ref. [25] is based on group method of data handling. This method initializes the missing values with mean imputation to generate the initial data set. This data set could be initialized with better imputation technique so that the number of iterations will be less. Missing value imputation by Ref. [25] takes more execution time. Moreover, Van Hulse and Khoshgoftaar [7] have only investigated the performance of five popular imputation techniques on noisy software measurement data set, but they have not proposed any technique to deal with missing values in the presence of noise.

In this paper, we have proposed an imputation technique based on DBSCAN mechanism. We hypothesize that the missing values can be calculated more accurately in the presence of noise using our proposed method. We have validated our hypothesis by experimenting on the Iris data set from standard repository [2] as well as 2D data set [10]. Moreover, we have examined the impact of noise by introducing different noise levels and different missing rates. Experimental results show that our algorithm performs better compared to K-means imputation [13] in the presence of noise.

3 Proposed Work

Our motivation for this study is based on the clustering algorithms that groups instances within a specified radius on the basis of distance principle such as the Euclidean distance. Density-based spatial clustering of applications with noise (DBSCAN) [4] proposed by Kriegel is a clustering technique that locates the cluster on the basis of density. This approach calculates the density in the data set by counting the number of points within a specified radius. It is relatively noise resistant and handle clusters of arbitrary shape as against K-means clustering algorithm.

We have applied density-based clustering method on whole data set in order to group the instances in cluster based on their distances. It also separates out the noisy instances, thus, minimizing their impact on MVI. This tends to better accuracy compared to the case when no elimination of noisy data is made prior to MVI. Later on, we make use of cluster information to find the nearest neighbors of missing value attributes instance. The average of missing attribute values of the nearest neighbors is used to impute missing value attributes.

3.1 Notions

In this section, we have used some assumptions to compute missing values and definitions used in clustering algorithm.

Let us consider a given set of N instances $P = (p_1, p_2, \ldots, p_N)$, where each instance has A set of attributes, we refer p_{ij} ($1 \le i \le N$ and $1 \le j \le A$) to indicate the value of attribute j in instance p_i. If instance p_i satisfies the condition $\{p_{ij} \ne ¢, 1 \le j \le A\}$, we can say that it is a complete instance. We call an instance as incomplete if $\{p_{ij} = ¢, 1 \le j \le A\}$, and thus, we can say that p_i has a missing value on attribute j. Let us consider set $R\{p_{ij} \ne ¢, 1 \le j \le R\}$ be the set of attributes whose value is present and call such attributes as reference attributes. Our main aim is to find the values of missing attribute for incomplete instances.

Let us consider following definitions [4] in the proposed approach:

- Eps: Maximum radius of the neighborhood
- MinPts: Minimum number of instances in an Eps neighborhood of that instance

Eps neighborhood of a point: The *Eps- neighborhood* of an instance p, denoted by $N_{\text{Eps}}(p)$, is defined by

$$N_{\text{Eps}}(p)\text{: } \{q \text{ belongs to } p \mid d(p_i, q_i) <= \text{Eps}\} \text{ where,}$$

- q_{ij} indicates the value of attribute j in instance q_i, and $1 \le i \le N$.
- $d(p_i, q_i)$ is evaluated using Eq. (1), if p_i and q_i are complete instances.
- $d(p_i, q_i)$ is evaluated using Eq. (2), if at least one of p_i and q_i is incomplete instance.

$$d(p_i, q_j) = \sqrt{\sum_{j=1}^{A} (p_{ij} - q_{ij})^2} \tag{1}$$

$$d(p_i, q_j) = \sqrt{\sum_{j=1}^{R} (p_{ij} - q_{ij})^2} \tag{2}$$

3.2 Choice of Eps and Minpoints

We have taken the value of MinPts as 4 for both the data sets at 0% noise level for experimentation. The value of Eps is computed as 0.2037 and 1.974 for Iris and 2D data set respectively [4, 14]. As the density-based clustering algorithm DBSCAN creates a cluster when it contains more than MinPts in the Eps neighborhood of a given instance, hence, the value of minpoints is set according to the noise level that we generated in the data set.

$$\text{MinPts} = \text{dataset size} * \text{noise level} + 1$$

i.e. for 6% noise level, MinPts = 150 * 0.06 + 1 = 10
i.e. for 12% noise level, MinPts = 150 * 0.12 + 1 = 19

3.3 Proposed Algorithm

This section describes the proposed imputation technique using density-based clustering algorithm because of its noise resistance. The proposed method takes two parameters, i.e. Eps and MinPts as inputs and consists of the following steps:
Step 1: Arbitrary choose an instance p from the data set.
Step 2: Retrieve all the instances that are density reachable from p with respect to Eps and MinPts. An instance p is density reachable from another instance q w.r.t Eps, MinPts if there is a chain of instances $p_1 \ldots p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density reachable from p_i. A point p is directly density reachable from a point q w.r.t Eps, MinPts if

$$p \text{ belongs to } N_{\text{Eps}}(q) \text{ and } N_{\text{Eps}}(q) \mid >= \text{MinPts}$$

Step 3: If p is an instance with at least MinPts instances within a radius "Eps-neighborhood", a cluster is formed.
Step 4: If p is on the border of the cluster, no instances are density reachable from p, and the next instance in the data set is visited.
Step 5: Continue the process until all of the instances are traversed.
Step 6: Identify the clusters after step 5. A cluster C with respect to Eps and MinPts is a non-empty subset of D satisfying the following conditions:
a) $\forall\, p, q$: if $p \in C$ and q is density reachable from p with respect to Eps and MinPts, then $q \in C$.
b) $\forall\, p, q \in C$: p is density connected to q with respect to Eps and MinPts

The instances that do not belong to any cluster are considered as noisy instances.
Step 7: Take all instances within the belonging cluster as nearest neighbors for each incomplete value instance.
Step 8: Use the averaged value of nearest neighbors for a missing attribute to fill in the missing attribute of incomplete instance.

4 Experimental Framework, Results, and Analysis

In this section, we have investigated the impact of noise on imputation techniques with a different missing rate to validate the efficiency of the proposed approach. Moreover, we have compared the proposed method with a popular K-mean imputation technique (KMI) [13].

4.1 Data Sets

Two data sets were used to validate the proposed approach. The first data set is Iris from the University of California at Irvine (UCI) machine learning repository [2] and was used in the experiments. This data set consists of 150 instances described by four attributes and one class attribute that depicts the class of a plant. Out of 150 instances, 50 instances belong to class Iris-setosa, next, 50 belong to class Iris-versicolour, and the last 50 instances belong to class Iris-virginica. The description of the four attributes is mentioned in the Table 1.

The second data set is Jain's (2D) data set [10] that consists of 373 instances. This data set is described by three attributes, namely, x, y, and a class. Both the attributes x and y have numerical values, and a class attribute has two categories, namely, "yes" and "no." Out of 373 instances, 97 instances belong to class "yes" and 276 instances belong to class "no."

Analysis has been carried out on four numeric attributes of Iris data set and two numeric attributes of 2D data set to validate the proposed method. The complete data set was chosen for the missing vale analysis. It is the ground truth that we can evaluate the performance if we are aware of the exact values of the attribute.

4.2 Experimental Design

We have taken two important factors, namely, noise and missing values in our experiments to evaluate the proposed method. The proposed method as well as KMI has been implemented in Matlab 7.0.1. Noise is present in the data set when the values of one or more attributes of an instance are corrupted or incorrect. The UCI data sets have been carefully examined by the domain experts, and they do not contain much noise [7]. We have manually injected noise to sepal length variable and x variable in Iris and 2D data sets, respectively. Randomly, we have introduced four levels of noise in one variable, i.e. such as 0%, 3%, 6%, and 12%. 0% indicates no noise, 3% shows that the value of sepal length variable is corrupted in 3% of the total instances and so on. The value of sepal length variable was changed to the opposite extreme of the univariate distribution. The noise generation process [12] in sepal length as well as x variable is as follows:

Assume that:
M is the attribute to be corrupted,
M_k is the value of the attribute M for the k^{th} instance of the data set to be corrupted,
max (M) is the maximum value for attribute M,
min (M) is the minimum value for attribute M,
Median (M) is median value for attribute M,
M'_k is the new value of attribute M after corruption,
$K = 60\% * \max(M)$,
Then,

- If $M_k < \text{Median}(M)$ then $M'_k = \max(M) + K$.
- If $M_k > \text{Median}(M)$ then $M'_k = \min(M) - K$
- If $M_k = \text{Median}(M)$ then $M_k = \text{random}(\min(M) - K, \max(M) + K)$

Table 1: Data set description.

Attribute name	Type of attribute
Sepal length	Numeric
Sepal width	Numeric
Petal length	Numeric
Petal width	Numeric
Class attribute	Nomoinal

Any instance is said to have missing values if the value of some attribute/attributes are not available. As we have chosen the data sets where all the attribute values were present, we have generated the missing data by removing the values randomly by eq. (3).

$$\text{Row} = \text{data set size} * \text{random}() \tag{3}$$

where, Row is the instance number of the data set, which can vary from 0 to 150 for Iris data set, and random is a function that generates the random number between 0 and 1. To control over missingness on sepal width variable, six levels of missing rate have been used, i.e. 2%, 4%, 6%, 8%, 10%, and 12%. In total, we have used $4 * 6 = 24$ combinations to validate our proposed imputation technique. For missingness, we have artificially deleted the data from the data set. Experiments were repeated for 2D data set as well.

4.3 Performance Measure

To evaluate the accuracy of imputation techniques on Iris data set, which belongs to life science domain, and 2D data set that belongs to shape data set [10] the root mean square error (RMSE) [15] is used as the performance measure. This measure is defined in eq. (4)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_{ij} - e_{ij})^2} \tag{4}$$

where a_{ij} is the actual value, e_{ij} is an estimated or imputed value of the j^{th} attribute of the i^{th} instance of the data set, and n is the total number of missing values. In our experiments, we have randomly removed the amount of data from the complete data set to generate the missing values and compared with their actual values to obtain the error value because it is the ground truth that we can evaluate the performance if we are aware of the exact values of the attribute.

4.4 Results and Discussion

We have performed the experiments on two data sets, namely, Iris and 2D. Figures 1 and 2 illustrate all the combinations of the different missing rates and noise levels in the Iris as well as the 2D data set, respectively, where the *x*-axis indicates the level of missing rate, and the *y*-axis represents the RMSE calculated by eq. (4). There are two curves in each part of Figures 1 and 2. The red curve represents the result of the K-means imputation [13], and the blue curve shows the result of the proposed method, namely, DBSCANI.

We have carried out numerous experiments by generating four levels of noise such as 0%, 3%, 6%, and 12% in one variable. For each level of noise, different levels of missing rates such as 2%, 4%, 6%, 8%, 10%, and 12% have been used to control the missingness in the data set. Moreover, we have conducted the experiments for KMI to compare with the proposed approach.

4.4.1 Discussion for Iris Data Set

The following are the observations from Figures 1 and 2:

- For DBSCANI at 0% noise level as shown in Figure 1A, at a small missing rate (2%), RMSE is more compared to the medium missing rate, i.e. 4%. For a higher missing rate (6%, 8%, 10%, and 12%), the error is more compared to the medium missing rate as well as the small missing rate. For KMI at 0% noise level, RMSE increases as we increase the missing rate from 2% to 12%.
- For DBSCANI at 3% noise level as shown in Figure 1B, increasing the missing rate from 2% to 12%, there is no significant change in RMSE. But KMI gives the highest imputation error at 4% compared to the other level of missing rate.

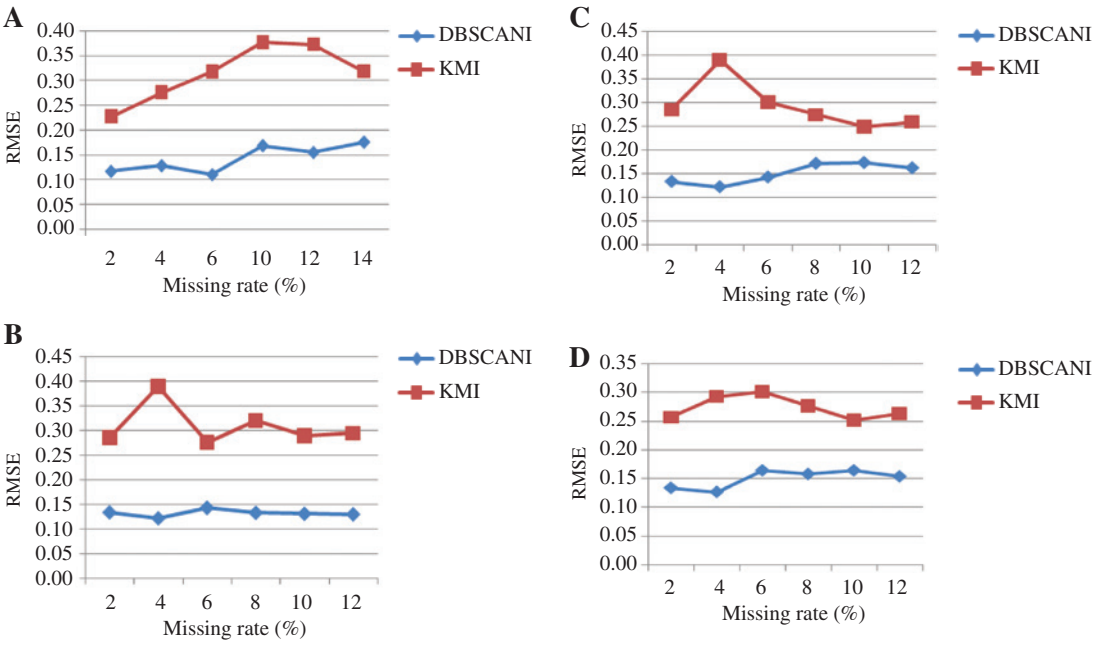

Figure 1: Experimental Results on Iris Data Set.
(A) Noise level = 0%. (B) Noise level = 3%. (C) Noise level = 6%. (D) Noise level = 12%.

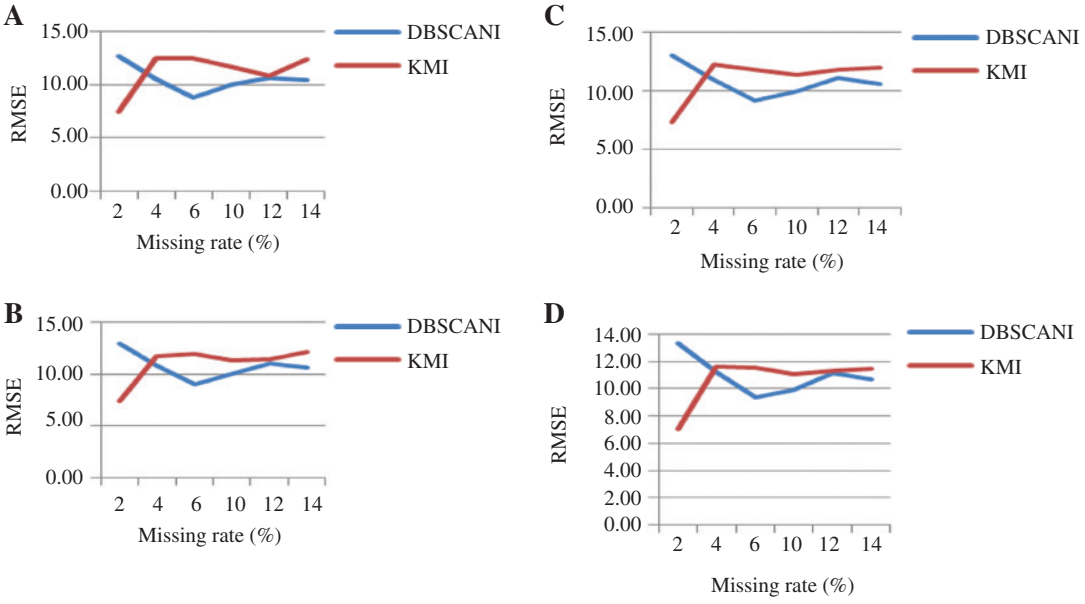

Figure 2: Experimental Results on the 2D Data Set.
(A) Noise level = 0%. (B) Noise level = 3%. (C) Noise level = 6%. (D) Noise level = 12%.

- At 6% of noise level as shown in Figure 1C, the impact of noise is significant, and imputation error increases at a higher missing rate (6%, 8%, 10%, and 12%). For KMI, the impact of noise is insignificant and the average imputation error is less compared to the 3% level of noise.
- At 12% of the noise level as shown in Figure 1D, DBSCANI performs the same compared to 6% of the noise level. KMI also performs better compared to the 6% level of noise.

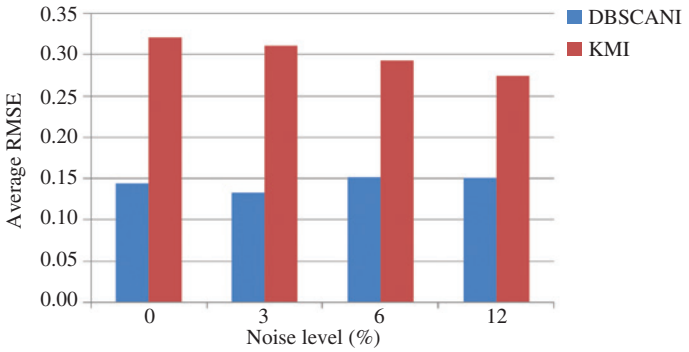

Figure 3: Average RMSE versus Noise Level for Iris Data Set.

– For DBSCANI, at a low noise level (3%) as shown in Figure 3, the impact of the noise is limited, and imputation errors are less. Sometimes, a small amount of noise (3%) even seems to improve the results as in Figures 1B,C and 3. However, when the level of noise is relatively high, the introduction of more noise (6% and 12%) will weaken the imputation results. For KMI, increasing the noise level decreases the imputation errors. But overall, the imputation error achieved by the proposed method is much less compared to KMI.

Therefore, we can conclude that DBSCANI performs better compared to KMI at different noise levels, i.e. 0%, 3%, 6%, and 12% as well as at a different missing rate.

4.4.2 Discussion for 2D Data Set

The following are the observations from Figures 2 and 4.

– For DBSCANI, at a small missing rate (2%), RMSE is more compared to other missing rates. The value of RMSE keeps on decreasing until 10% of the missing rate. At 12% of the missing rate, RMSE is a little higher compared to 4%, 6%, and 10% missing rate. For the highest missing rate, RMSE is further reduced. For KMI, RMSE is less at a lower missing rate (2%) at all noise levels, but it is drastically increased for all other missing rates.
– For DBSCANI, at a low noise level (3%) as shown in Figure 4, the impact of noise is limited, and imputation errors are less. As we increase the noise in the data set, imputation errors are also increased. For KMI, increasing the noise level decreases the imputation errors. But overall, the imputation error achieved by the proposed method is much lesser compared to KMI.

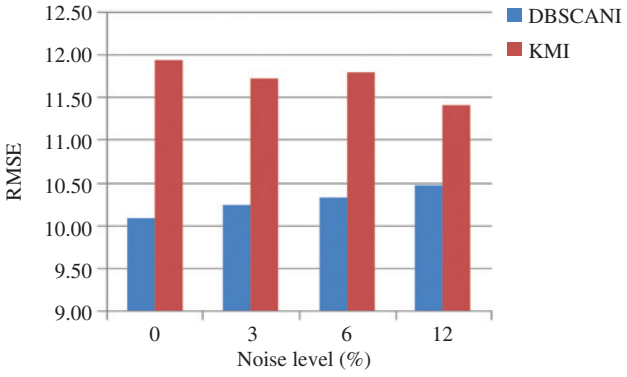

Figure 4: Average RMSE versus Noise Level for the 2D Data Set.

Therefore, we can conclude that DBSCANI performs better compared to KMI at different noise levels, i.e. 0%, 3%, 6%, and 12% as well as at different missing rates.

5 Conclusion

Missing value imputation has been a key area in data mining research. A great deal of missing value techniques have been discussed in the literature. The impact of noise has been neglected in most of the methods. Popular clustering methods such as K-means, fuzzy K-means, etc., have been used for MV imputation. This paper propounds a new technique that uses the DBSCAN clustering algorithm instead of the K-mean clustering to deal with the missing value in the presence of noise. We have systematically conducted experiments to analyze the impact of noise with the given missing rate. Moreover, we have illustrated our approach with a different missing rate at various noise levels. We conclude that DBSCANI performs better compared to KMI in a noisy as well as in a non-noisy environment. In the future, we would investigate the proposed approach with the noise as well as the missing values distributed throughout the data set.

Bibliography

[1] R. E. Abdel-Aal, GMDH-based feature ranking and selection for improved classification of medical data, *J. Biomed. Inform.* **38** (2005), 456–468.
[2] A. Asuncion and D. J. Newman, *UCI machine learning repository*, University of California, School of Information and Computer Science, Irvine, CA, 2007.
[3] A. P. Dempster, N. M. Laird and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *J. RS Soc. B* **39** (1977), 1–38.
[4] M. Ester, H. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining*, Portland, (1996), 226–231.
[5] O. Harel and X. H. Zhou, Multiple imputation: review of theory, implementation and software, *Stat. Med.* **26** (2007), 057–3077.
[6] R. J. Hathaway and J. C. Bezdek, Clustering incomplete relational data using the non-Euclidean relational fuzzy c-means algorithm, *Pattern Recognit. Lett.* **23** (2007), 151–160.
[7] J. V. Hulse and T. M. Khoshgoftaar, A comprehensive empirical evaluation of missing value imputation in noisy software measurement data, *J. Syst. Softw.* **81** (2008), 691–708.
[8] A. G. Ivakhnenko, The group method of data handling-a rival of the method of stochastic approximation, *Soviet Automatic Control* **1–3** (1968), 43–55.
[9] A. G. Ivakhnenko and Y. L. Kocherga, Theory of two-level GMDH algorithms for long-range quantitative prediction, *Soviet Automatic Control* **16**, (1983), 7–12.
[10] A. Jain and M. Law, Data clustering: a user's dilemma, *Lecture Notes in Computer Science* **3776** (2005), 1–10.
[11] M. Kamber and Jiawei Han, *Data mining concepts and techniques*, 2nd ed, Morgan Kaufman, Elsevier, 2006.
[12] T. Khoshgoftaar and V. Hulse, Empirical case studies in attribute noise detection, *IEEE Trans. Syst. Man Cybern.* **39** (2009), 379–388.
[13] J. Luengo, S. García and F. Herrera, On the choice of the best imputation methods for missing values considering three groups of classification methods, *J. Knowl. Inf. Syst.* **32** (2011), 77–108.
[14] M. Daszykowski, B. Walczak and D. L. Massart, *Looking for Natural Patterns in Data*. Part 1: Density Based Approach, Chemometrics and intelligent System, **56** (2001), 83–92.
[15] B. M. Patil, R. C. Joshi and D. Toshniwal, *Missing value imputation based on K-mean clustering with weighted distance*, IC3 Noida, India, pp. 600–609, Springer, 2010.
[16] V. Puig, M. Mrugalski, A. Ingimundarson, J. Quevedo, M. Witczak and J. Korbicz A GMDH neural network-based approach to passive robust fault detection using a constraint satisfaction backward test, *Eng. Appl. Artif. Intell.* **20** (2007), 886–897.
[17] A. Purwar and S. Singh, Empirical evaluation of algorithms to impute missing values for financial dataset, in: *International Conference on Issues and Challenges in Intelligent Computing Techniques*, Ghaziabad, 2014.
[18] J. L. Schafer, Multiple imputation: a primer, *Stat. Methods Med. Res.* **8** (1999), 3–15.
[19] S. Singh, H. Mittal, and A. Purwar, Prediction of investment patterns using data mining techniques, *Int. J. Comput. Commun. Eng.* **3** (2014), 145–148.
[20] P. N. Tan, M. Steinbach and V. Kumar, *Introduction to data mining*, Addison-Wesley, Boston, 2005.

[21] J. Tian, B. Yu, D. Yu and S. Ma, Clustering-based multiple imputation via Gray relational analysis for missing data and its application to aerospace field, *Sci. World J.* 2013.
[22] X. Wu and X. Zhu. Mining with noise knowledge: error-aware data mining, *IEEE Trans. Syst. Man Cybern. C* **38** (2008), 917–932.
[23] X. Zhu and X. Wu, Class noise vs. attribute noise: a quantitative study, Artif. Intell. Rev. **22** (2004), 177–210.
[24] X. Zhu, S. Zhang, Z. Jin, Z. Zhang and Z. Zu, Missing value estimation for mixed-attribute data sets, *IEEE Trans. Knowl. Data Eng.* **23** (2011), 110–121.
[25] B. Zhu, C. He and P. Liatsis, A robust missing value imputation method for noisy data, *J. Appl. Intell.* **36** (2012), 1–74.