

DATA MINING TECHNIQUES FOR MISSING VALUE IMPUTATION

Marghny H. Mohamed

Faculty of computers and Information, Assiut University, Egypt
marghny@aun.edu.eg

Abdel-Rahiem A. Hashem

Faculty of Science Assiut University, Egypt
hashem_aer2@yahoo.com

M. M. AbdelSamea

Faculty of Science Assiut University, Egypt
mm_abdelSamea@yahoo.com

(Received June 5, 2010 Accepted July 7, 2010)

Imputation is a class of procedures that aims to fill the values which are missed with estimated ones. These methods involve replacing missing values with estimated ones based on some information available in the data set. K-means has been successful in finding missing values for several data sets available such as Bupa, Breast Cancer, Pima, etc. In this paper, we introduce an efficient imputation methods based K-means to treat missing data. Our proposed methods give higher accuracy than the one on given by classical K-means. Experimental results hold on a variety class of data sets.

KEYWORDS: *Imputation, Clustering, K-mean*

1. INTRODUCTION

No quality data, no quality mining results [1]. Data quality is a major concern in Data mining and other correlated area such as Machine learning. Data mining refers to extracting knowledge from large amounts of data. One relevant problem in data quality is the presence of missing data. It is occurred in the phase of data collection [2]. Types of missing data: according to Little and Rubin, mechanisms that lead to missing data can be categorized into three types [3].

1. Missing completely at random (MCAR): the absence of a data element is not associated with any other value in the data set, observed or missing [4]. In other words, when the distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data [5].
2. Missing at random (MAR): this is less restrictive assumption than MCAR; it indicates that the absence of a data element depends only on the observed values in the data set, not on missing ones [4]. In other words, when the distribution of an example having a missing value for an attribute depends on the observed data, but does not depend on the missing data [5].
3. Not missing at random (NMAR): the condition is the negative of MAR. The absence of a data element reflects its probable (missing) data value [4]. In

other words, when the distribution of an example having a missing value for an attribute depends on the missing value [5].

Missing data treatment methods can be divided into three categories [2], [5].

1. Ignoring and discarding: There are two main ways to discard data with missing values. The first, complete case analysis. This method consists of discarding all instances with missing data. The second, discarding instances and/or attributes. This method consists of determining the extent of missing data on each instance and attribute and deleting the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with high degree of missing values [2].
2. Parameter estimation: Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm [6] can handle parameter estimation in the presence of missing data [2].
3. Imputation: Imputation [7], [8] is a class of procedures that aims to fill in the missing values with estimated ones [2].

2. IMPUTATION METHODS

Imputation methods [9] involve replacing missing values with estimated ones based on some information available in the data set. There are many options varying from naive methods like mean or mode imputation [10] to some more robust methods based on relationships among attributes.

Methods available for creating complete data matrices can be divided into two main categories: single imputation and multiple imputation methods. Single imputation methods fill in one value for each missing one; they have many appealing features, because standard complete-data methods can be applied directly and because imputation need to be carried out only once Multiple imputation methods generate multiple simulated values for each missing value, in order to reflect the uncertainty attached to missing data [11].

Mean and mode imputation (Mimpute). 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. Replacing all missing records with a single value distorts the input data distribution [12], [13], [14].

Hot deck imputation (HDimpute) and cold deck imputation (CDimpute). Given an incomplete pattern, HDimpute replaces the missing data with the values from the input vector that is closest in terms of the attributes that are known in both patterns [14], [15]. Unlike Mimpute, this method attempts to preserve the distribution by substituting different observed values for each missing item [12]. Another possibility is the CDimpute method which is similar to hot deck but the data source must be other than the current data set. For example, in a survey context, the external source can be a previous realization of the same survey [13].

Prediction models. These methods consist of creating a predictive model to estimate values that will substitute the missing data [14], [11]. The incomplete attribute with missing data is used as target, and the remaining attributes are used as inputs for

the model. An important argument in favour of this approach is that, frequently, attributes have relationships (correlations) among themselves. In this way, those correlations can be used to create a predictive model for classification or regression. The requirement for correlation among the attributes can be also a draw back in some situations. If there are no relationships among the incomplete feature and the remaining variables, then the model will not be precise to impute values for the missing ones. Its main disadvantage is that when missing items appear in many combinations of attributes in a high- dimensional problem, a huge number of prediction models has to be designed, i.e., one model per combination of incomplete attributes.

In this paper, we enhance an imputation method based k-mean in a several way by enhancement the way of imputation and gives an efficient accuracy compared with an imputation method based k-mean, which proved to be succeed in missing value imputation than other statistical approaches [2]

3. CLASSICAL K-MEAN

The k-means algorithm proceeds as follows. First, it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Classical K-mean algorithm describes as following:

1. **Initialize K centers ($w_1; w_2; \dots; w_k$) such that $w_j = x_i$, $j \in \{1; 2; \dots; k\}$, $i \in \{1; 2; \dots; n\}$, n number of objects each cluster C_j is associated with center w_j .**
2. **For each input vector x_i , where $i \in \{1; 2; \dots; n\}$, assign x_i to the cluster C_j with nearest center w_j**
3. **for each cluster C_j , Update the center w_j to be the centroid of all samples currently in C_j .**
4. **Repeat step 2,3 until the centroids do not change.**

4. MISSING VALUE IMPUTATION BASED CLASSICAL K-MEAN

After building the clusters using K-mean process and obtaining each clusters K centroids, we impute the missing values with corresponds prototypes from the most similar k-centroid. The Classic Imputation algorithm (CI) describes as:

1. **Divide dataset S into Complete-valued dataset S' , and Missing-valued dataset S^* .**
2. **Apply classical k-mean on complete dataset S' until convergence and obtain w_j centers, $j \in \{1; 2; \dots; k\}$**
3. **For each instance x_i containing missing value, where $x_i \in S^*$. Compute distance between centroid C_j and instance x_i containing missing value.**
4. **Impute missing-value in x_i from its corresponding closest centroid w_j .**

5. PROPOSED MODIFICATION OF CLASSIC IMPUTATION METHOD

When the missing values in the selected sample are exceeding the number of the available ones, this implies that the measured distance will be in $(n-p)$ space, which means inefficient measured distance. that's why we will improve the missing values imputation by modifying the steps to obtain the measured distance. When we are getting the first centroids from the clustering process, we initialize missing values by imputing from prototypes of these centroids, so the distance measure in the next step becomes in n dimension and in each new clustering process, imputation will be achieved by measuring the closest distance between whole sample and new centroids. The Modification of Classic Imputation algorithm (MCI) describes as, see Figure 1.

1. Divide dataset S into Complete-valued dataset S^c , and Missing-valued dataset S^* .
2. Start K-mean algorithm on S^c , while clusters optimized, for each computed centroid $w_j, j \in \{1; 2; \dots; k\}$ and missing-value instance x_i , where $x_i \in S^*$. Compute distance between centroid w_j and missing-value instance x_i .
3. Impute missing-value in x_i from its corresponding closest centroid w_j .
4. Repeat step 2 and 3 until k-mean convergence.

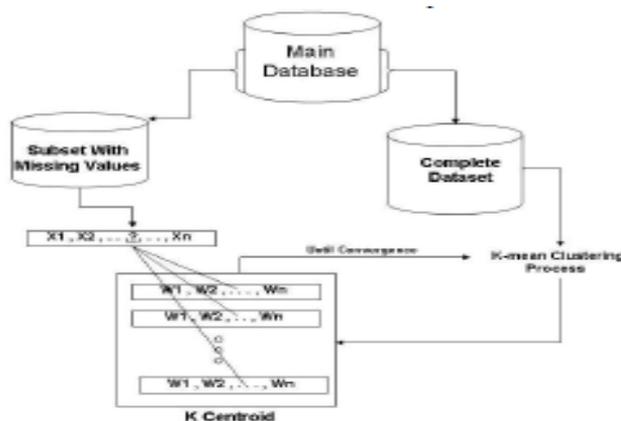


Figure 1. A Modification of classic Imputation based kmean.

6. ENHANCEMENT OF MODIFICATION OF CLASSIC IMPUTATION METHOD

In each clustering process each sample gets imputed from the centroid of its closest cluster, we count the number of times the sample has been imputed from a particular cluste. The largest number of times a sample gets assigned to a particular cluster means that it belongs to this cluster, which will result on the imputation of the values of the last cluster's centroid of the most visited cluster to the sample. Enhancement of Modification of Classic Imputation algorithm (EMCI) describes as follow:

1. Divide dataset S into Complete-valued dataset S' , and Missing-valued dataset S^* .
2. Initialize class counter CC_j for each missing-value instance, where $j \in \{1; 2; \dots; k\}$.
3. Start K-mean algorithm on S^* , while clusters optimized, for each computed centroid w_j , $j \in \{1; 2; \dots; k\}$ and missing-value instance x_i , where $x_i \in S^*$. Compute distance between centroid w_j and missing-value instance x_i .
4. Impute missing-value in x_i from it's corresponding closest centroid w_j and increment it's corresponding closest center cc_j .
5. Repeat step 3 and 4 until k-mean convergence.
6. For each missing-value instance x_i , where $x_i \in S^*$, Choose the maximum class counter and impute missing-value in x_i with it is corresponding prototype centroid.

7. Experimental results

We choose four real-world data sets from the UCI Machine Learning Repository [3] and compare the three missing value strategies discussed earlier. These data sets are chosen because they have at least some discrete attributes, multi class, and a good number of examples and we will select values from original data sets to be missing to simulate different situations with missing values. To simulate missing values in data sets, we randomly select certain percentages (2 percent, 4 percent, 6 percent, 10 percent, 20 percent and 40 percent) of attribute values in the whole data set to be missing and those missing values are distributed into each attribute proportional to its cost as more expensive attributes usually have more missing values.

Table 1. Data Sets Used in the Experiments

	No. of attributes	No. of examples
Iris	4	150
Ecoli	7	336
Bupa	6	345
Pima Indian	8	768

This study shows the performance of three imputation methods based k-mean; Classic Imputation (CI), Modification of Classic Imputation (MCI) and Enhancement of Modification of Classic Imputation (EMCI). Each graph compares the performance of all methods with different level of missing values for different clusters of K-Mean. We are use for an accuracy the mean square errors which give from error = $(R - I)^2/N$ where R is real value, I is Imputed value and N is number of missing values.

In our experimental, all figures illustrate the mean square error comparison for the three imputation method describes in previous sections, while all tables illustrate the sum of square errors comparison for simplicity of showing the difference between three methods.

Table 2 illustrates an error comparison between an imputation methods based k-mean, CI, MCI and EMCI in different missing instance percentage at several cluster number for Bupa dataset.

Table 2. A sum of square error comparison of three imputation methods in the Bupa data set.

Miss.(%)	Cluster No	Imputation approaches based K-mean		
		CI	MCI	EMCI
2	3	4.777186	3.186759	0.299139
	4	3.581818	2.791975	0.309811
	5	0.907024	0.907024	0.684761
	6	1.116783	1.116783	1.116783
	7	0.95998	0.95998	0.95998
4	3	4.383957	3.588674	0.609776
	4	4.39595	2.953905	0.489464
	5	1.481603	1.481604	1.183803
	6	1.669874	1.669874	1.669874
	7	1.527588	1.527589	1.527589
6	3	7.765477	6.71576	4.331328
	4	6.907715	5.638264	3.821443
	5	3.641498	3.641498	3.763348
	6	7.375067	7.375067	7.375067
	7	3.671032	3.671032	3.671032
10	3	12.01079	11.09118	10.11517
	4	11.92479	10.60073	8.489751
	5	8.332617	8.33262	8.420806
	6	14.09595	14.09595	14.09595
	7	8.28245	8.282451	8.282451
20	3	12.76048	11.13619	6.747862
	4	11.67825	10.88445	5.91155
	5	10.84042	10.84041	10.16937
	6	23.74567	23.74565	23.74565
	7	11.98075	11.98074	11.98074
40	3	18.34416	16.73777	15.22232
	4	21.26251	20.46837	13.66378
	5	22.84709	22.84707	19.59215
	6	47.41088	47.41081	47.41081
	7	14.56487	14.56487	14.56487

From table 2 we can notice that the difference between EMCI algorithm and other two methods is clear strongly for little clusters in this data set, Bupa, see Fig. 2.

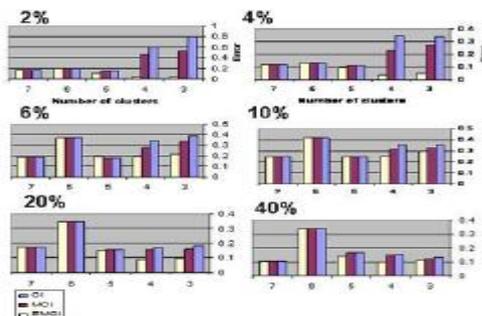


Figure 2. Mean square error comparisons of three imputation methods in the Bupa data set.

Table 3 illustrates an error comparison between an imputation methods based k-mean, CI, MCI and EMCI in different missing instance percentage at several cluster number for Pima Indian dataset.

Table 3. A sum of square error comparison of three imputation methods in the Pima Indian data set.

Miss.(%)	Cluster No	Imputation approaches based K-mean		
		CI	MCI	EMCI
2	3	14.37554	13.25002	8.890231
	4	13.85666	12.84985	6.06536
	5	13.37925	12.84137	5.971876
	6	11.82608	11.45384	5.877658
	7	5.825113	5.825112	5.916673
	8	7.825108	7.825106	7.825106
4	3	18.03262	17.36147	13.55296
	4	18.93389	17.38975	12.54147
	5	18.04683	17.85205	12.71843
	6	16.84882	16.74036	11.74633
	7	11.36014	11.36014	11.91438
	8	11.42276	11.42276	11.42276
6	3	24.44564	23.20973	17.91235
	4	22.63036	21.62001	16.1246
	5	20.38086	20.01029	15.99459
	6	20.81671	20.33147	14.35323
	7	14.01545	14.01545	14.26166
	8	14.28277	14.28277	14.28277
10	3	38.09534	36.341	33.59278
	4	33.94895	33.3278	29.87464
	5	33.55535	33.02003	30.1552
	6	32.1873	31.36938	27.59835
	7	26.59483	26.59483	27.42991
	8	26.49616	26.49616	26.49616

20	3	64.64467	62.69846	64.57974
	4	64.20623	63.53032	64.39197
	5	81.72205	81.28112	70.34858
	6	64.43491	63.68247	59.91234
	7	57.20836	57.20828	58.84161
	8	57.41436	57.41428	57.41428
40	3	119.333	117.3868	127.27
	4	119.7724	119.106	127.9656
	5	118.6994	118.2354	127.2739
	6	120.2016	119.4625	117.9062
	7	113.2273	113.2272	115.7847
	8	145.5723	145.5722	145.5722

From Table 3 we can notice that EMCI is still better than other two methods. For higher level of missing 20 and 40 percentage for little clusters 3 and 4 the MCI is better, see Fig. 3.

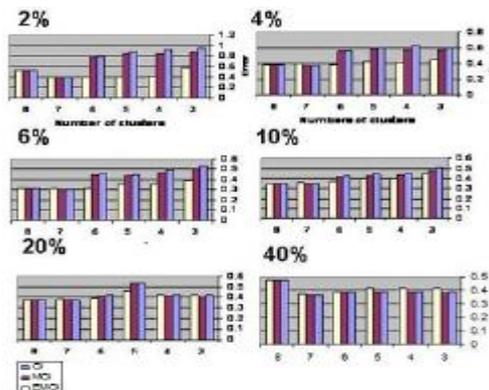


Figure 3. A mean square error comparison of three imputation methods in the Pima Indian data set.

Table 4 illustrates an error comparison between an imputation methods based k-mean, CI, MCI and EMCI in different missing instance percentage at several cluster number for Ecoli dataset.

Table 4. A sum of square error comparison of three imputation methods in the Ecoli data set.

Miss.(%)	Cluster No	Imputation approaches based K-mean		
		CI	MCI	EMCI
2	3	5.41087	4.531235	4.075765
	4	9.091621	8.014148	3.787097
	5	4.477868	4.462628	3.313772
	6	2.875017	2.875017	3.065405
	7	2.676449	2.676449	2.676449

4	3	6.562714	5.990744	6.461079
	4	10.32932	8.963055	6.190022
	5	6.512977	6.551205	5.492534
	6	4.635415	4.635415	5.314992
	7	4.161032	4.161032	4.161032
6	3	7.599198	7.035617	9.206245
	4	11.0807	9.722828	8.954317
	5	7.268047	7.306274	7.490683
	6	5.395188	5.395188	5.761791
	7	5.368557	5.368557	5.368557
10	3	9.096494	8.35237	10.79339
	4	12.58382	11.04685	9.157546
	5	10.30394	10.36206	8.858151
	6	6.638017	6.638016	7.07966
	7	6.767209	6.767208	6.767208
20	3	14.13585	13.38966	22.07389
	4	17.63459	16.09575	19.19012
	5	16.69709	16.76461	18.83994
	6	11.52075	11.52075	13.28862
	7	15.41178	15.41178	15.41178

From table 4 we can notice that the two methods MCI and EMCI is better than CI and in many cases the MCI is better than the EMCI but the EMCI is better in other cases, See figure 4.

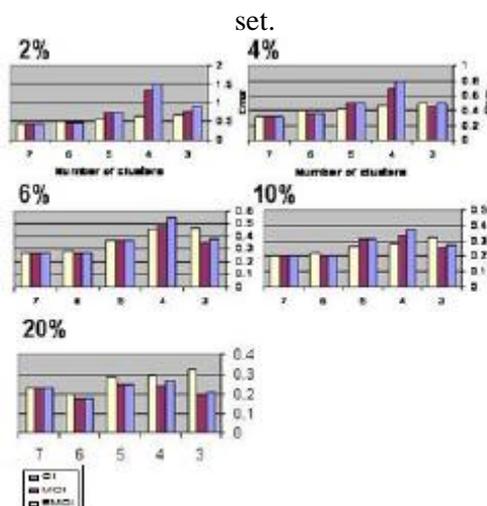


Figure 4. A mean square error comparison of three imputation methods in the Eoli data

Table 5 illustrates an error comparison between an imputation methods based k-mean, CI, MCI and EMCI in different missing instance percentage at several cluster number for Iris dataset.

Table 5. A sum of square error comparison of three imputation methods in the Iris data set.

Miss.(%)	Cluster No	Imputation approaches based K-mean		
		CI	MCI	EMCI
2	3	0.447209	0.447209	0.447209
	4	0.420188	0.420188	0.420188
	5	0.420188	0.420188	0.420188
	6	0.457572	0.457572	0.457572
	7	0.536267	0.536267	0.536267
4	3	1.018319	1.018319	1.06553
	4	0.790119	0.79012	0.79012
	5	0.790119	0.79012	0.79012
	6	1.105146	1.105146	1.105146
	7	1.278171	1.278171	1.278171
6	3	1.438984	1.438984	0.893171
	4	1.066982	1.066982	1.066982
	5	1.066982	1.066982	1.066982
	6	1.588675	1.588675	1.588675
	7	1.926627	1.926627	1.926627
10	3	2.252538	2.252538	1.755318
	4	1.629589	1.629589	1.629589
	5	1.629589	1.629589	1.629589
	6	2.499124	2.499124	2.499124
	7	3.06347	3.063469	3.063469
20	3	4.289041	4.28904	4.832908
	4	3.045854	3.045854	3.045854
	5	3.045854	3.045854	3.045854
	6	4.535299	4.535299	4.535299
	7	5.684607	5.684606	5.684606
40	3	6.942563	6.942573	7.727729
	4	9.3234	9.323404	9.323404
	5	4.254536	4.254535	4.254535
	6	6.270786	6.270789	6.270789
	7	7.323775	7.323776	7.323776

From table 5 we can notice that the three algorithms are same in most cases. EMCI is better for level of missing value 6 and 10 percentage for little clusters as 3, see Fig. 5.

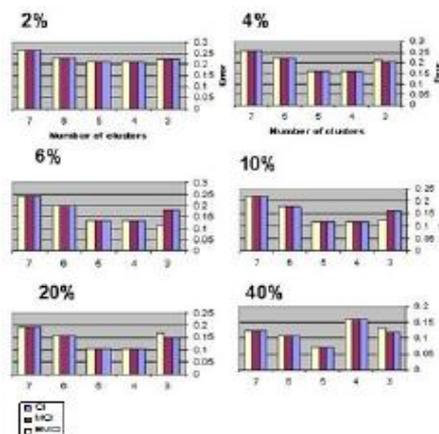


Figure 5. A mean square error comparison of three imputation methods in the Iris data set.

8. CONCLUSIONS

Missing data is a usual drawback in many real-world applications. A classical solution is imputation, i.e., to estimate and to fill in the unknown values using available data. This work analyzes the behavior of three imputation methods based on k-mean; a classic imputation (CI), a modification of classic imputation (MCI) and enhancement of modification of classic imputation (EMCI). The first method (CI) is used and gives higher accuracy than Mean, Mode, Median and c4.5 on dataset such as Bupa, Pima Indian, e.t. Our proposed methods; (MCI) and (EMCI) is better than the classic (CI).

In most cases when the number of clusters is less, the performance of EMCI is better than the two other methods and MCI is better than CI. When the number of clusters increases, the three algorithms are the same.

REFERENCES

- [1] Jiawei, H. and K. Micheline, 2006. Data mining Concept and Techniques. 2nd Edn Morgan Kaufmann Publishers. ISBN: 1-55860-901-6.
- [2] B. Mehala, K. Vivekanandan and P. Ranjit Jeba Thangaiah, An Analysis on K-Means Algorithm as an Imputation Method to Deal with Missing Values. Asian Journal of Information Technology 7 (9): 434-441, 2008.
- [3] K. Lakshminarayan, S. A. Harp, T. Samad, Imputation of missing data in industrial database, Apple. Intell. 11 (1999) 259-275.
- [4] Jau-Huei Lin, Peter J. Haug, Exploiting missing clinical data in Bayesian network modeling for predicting medical problems Journal of Biomedical Informatics 41 (2008) 1-14.
- [5] Alireza farhangfar, Lukase Kurgan and Jennifer Dy, Impact of imputation of missing values on classification error for discrete data. Pattern Recognition 41 (2008) 3692-3705.
- [6] Dempster, A.P., R.J. Laird and D.B. Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm (with Discussion). I. R. Stat. Soc, B39: 1-38. <http://www.jstor.org/pss/2984875>.

- [7] Daqian, G. and G. Yang, 2005. Incremental gradient descent imputation method for missing data in learning classifier systems. GECCO, ACM, Washington, DC, USA, pp: 72-73. DOI: <http://doi.acrn.org/10.1145/1102256.1102270>, <http://portal.acm.org/citation.cfm?id=1102270> and CFID=6188267 and CFTOKEN=42900381
- [8] Fulufhelo, V., Nelwamondo and M Tshlidzi, 2007. Rough sets computations to impute missing data. Comput. Vision and Pattern Recog., 1: 1-19.
- [9] Musil, C.M., C.B. Wamer, P.K. Yobas and S.L. Jones, 2002. A comparison of imputation techniques for handling missing data. Western J. Nus. Res., 24 (5).
- [10] Cristian P., D. Alain, P. Monique and K. Tahar, 2005. Tools for statistical analysis with missing data: Application to a large medical database. ENMI, pp: 181-186 <http://lwm.magic5.unile.it/PapDoc>
- [11] Josef L. Schaffer. Multiple Imputation for multivariate Missing data problems: A Data Analyst's Perspective.
- [12] Pedro J.Garc Laencina, Jose Luis Sancho-Gomez, Anbal R.Figueiras-Vidal, Michel Verleysen, K nearest neighbours with mutual information for simultaneous Classification and missing data imputation. Neurocomputing 72(2009)1483-1493.
- [13] P.D.Allison, Missing data, Sage University Papers Series on Quantitative Applications in the Social Sciences, Thousand Oaks, California, USA, 2001.
- [14] R. J. A. Little, D. B. Rubin, Statistical Analysis with Missing Data, second edition, Wiley, NJ, USA, 2002.
- [15] G. Sande, Hot Deck Imputation Procedures, Incomplete data in Sample Surveys, vol. 3, Academic Press, New York, 1983.

استخدام طرق تنقيب البيانات في إحلال القيم المفقودة

يعتمد التحليل الإحصائي والقرارات والنتائج المترتبة عليه لمجموعة من البيانات الخاصة بظاهرة ما على كون البيانات كاملة، فإذا وجد بها نقص في بعض متغيراتها في بعض الحالات (الأمثلة) فإن ذلك يسبب نتائج متحيزة *baised results* يعني متحيزة للبيانات الموجودة وبالتالي لن تكون هذه النتائج صحيحة بالقدر الذي لو كانت البيانات موضع التحليل كاملة.

في هذا البحث نحاول استخدام أحد تقنيات طرق تنقيب البيانات *Data mining techniques* - وهو الخوارزم *Kmean* والمستخدم أساساً في عمل تصنيف غير موجه بنوع الفئة أو المجموعة وهي العملية المسماة *Clustering* - في إحلال القيم المفقودة *Imputation* بقيم تقريبية تعتمد على تقسيم جزء البيانات الكاملة أولاً إلى مجموعات متشابهة الخصائص *clusters* ثم محاولة المفقودة بقيم متوسط هذه المجموعة.

ثم بتعديل طريقة الإحلال *Imputation* المعتمدة على الخوارزم *Kmean* ثم بتحسين هذا التعديل تم الحصول على نتائج أفضل في معظم الحالات.