Solving Missing Values : A Case Study

Mansoor Habeeb
Dept of Mathematis/Faculty of Maths and CS, University of Kufa
Mansoor.Habeeb@uokufa.edu.iq
Kadhim Aljanabi
Dept of Computer Science/Faculty of Maths and CS, University of Kufa
Kadhim.aljanabi@uokufa.edu.iq
Nawras Riyadh Neamah
Dept of Mathematical Faculty of Maths and CS, University of Kufa
Nawras.8888@yahoo.com

Abstract:

One of the most important issues in information theory related to data in both Database and Data Warehouse is the missing values (unknown, not available and required). This represents a great challenge to the analysis process. Features or data attributes (fields or columns in relational DB) in data repositories represent the core of any analytical process in OLAP(**O**n **L**ine **A**nalytical **P**rocessing)and OLTP(**O**n **L**ine **T**ransaction **P**rocessing). These attributes are required to be studied and processed. Many papers were published to solve such problem in different goals and algorithms. However, the aim of this research proposal is to improve the algorithms applied to these topics to insure data consistency, correctness, completeness, and time and space complexity. Different algorithms and techniques were applied on more than 20000 records collected from different hospitals and clinics around Iraq to study the effectiveness of the proposed algorithms including Most Common Value, overall average, and classification.

Research Methodology:

1. Data Collection
2. Data Preprocessing
3. Missing values estimation
4. Algorithms improvement
5. Result analysis

Keywords: DB, DW, OLTP, OLAP, Missing Values

الملخص:

تعتبر القيم المفقودة من أهم القضايا في نظرية المعلومات المتعلقة بالبيانات في كل من قاعدة بيانات ومخازن البيانات (وهي تشير الى القيم غير المعروفة، وليست متاحة ولكنها مطلوبة) وهذا الموضوع يمثل تحديا كبيرا لعملية تحليل البيانات. ميزات أو سمات البيانات (الحقول او الاعمدة في انظمة قواعد البيانات) تمثل جوهر أي عملية تحليلية في OLAP(انظمة المعالجة التحليلية المباشرة) و OLTP(انظمة معالجة الاجراءات المباشرة). وهذه الصفات تتطلب الدراسة والمعالجة. وقد تم معالجة هذه المشكلة في العديد من الاوراق البحثية باستخدام خوارزميات ومنهجيات مختلفة، وتهدف هذه الورقة البحثية الى تحسين وتطوير الخوارزميات المطبقة على هذه المواضيع لضمان اتساق البيانات وصحتها واكتمالها وتطويرها في كل من معياري الوقت والسعة المطلوبة لانجاز الخوارزمية. وتم تطبيق هذه الخوارزميات على اكثر من 20000 سجل تم جمعها من مستشفيات وعيادات مختلفة في العراق. وقد تم تطبيق الخوارزميات التالية:

- القيمة الأكثر شيوعا ضمن البيانات
- المعدل العام ومعدل الفئات المختلفة

- التصنيف

منهجية البحث

- تجميع البيانات
- المعالجة الاولية للبيانات
- تعويض البيانات المفقودة
- تطبيق الخوارزميات والمنهجيات
- تحليل النتائج

1. INTRODUCTION

The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. Data storage became easier as the availability of large amounts of computing power at low cost, the cost of processing power and storage is falling, made up data storage cheap.

Having concentrated so much attention on the accumulation of data the problem was what to do with this valuable resource? It was recognized that information is at the heart of business operations and that decision-makers could make use of the data stored to gain valuable insight into the business. Database Management systems gave access to the data stored but this was only a small part of what could be gained from the data. Traditional on-line transaction processing systems (OLTPs) are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return. Analyzing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business. This is where Data Mining or Knowledge Discovery in Databases (KDD) has obvious benefits for any enterprise [1, 2, 3].

The term data mining has been stretched beyond its limits to apply to any form of data analysis. Some of the numerous definitions of Data Mining, or Knowledge Discovery in Databases are:

Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency net works, analyzing changes, and detecting anomalies [2, 4].

Knowledge extraction consists of the following steps:

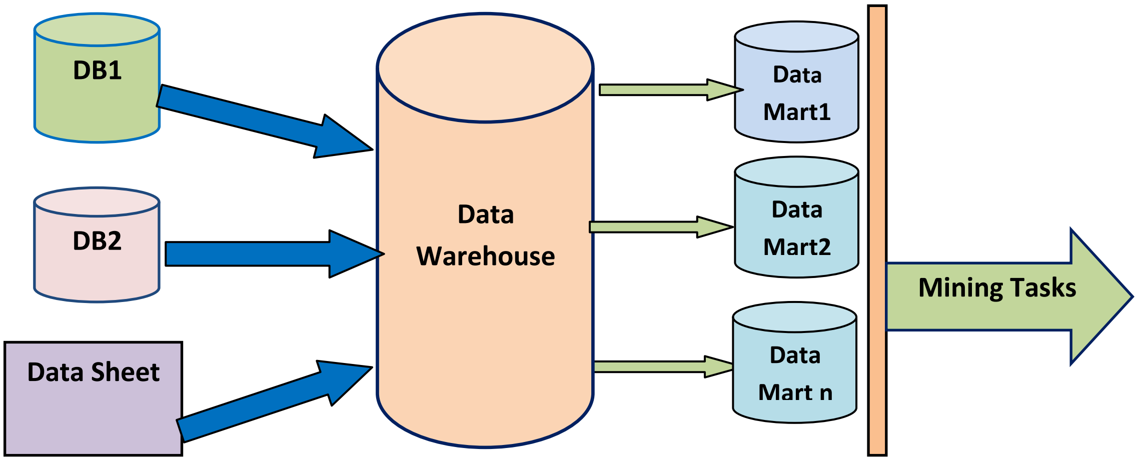
1. Data Selection
2. Data Preprocessing
3. Transformation
4. Data mining
5. Interpretation and evaluation

DM tasks can be summarized into the following categories:

1. Classification
2. Association
3. Clustering
4. Trends

5. Prediction
6. Link Analysis.

   Each of them has its own techniques, algorithms, and applications. DM usually applied on huge data repository known as Data Warehouse that may be divided into many sub warehouses known as data marts, see figure(1).



Data from different source Data Repository (DW) Data Marts Mining algorithms and reporting

Figure (1): Data Warehouse Architecture.

2. Data Preprocessing

Real world data usually have the following drawbacks: Incompleteness, Noisy and Inconsistence. So, these data need to be preprocessed to get the data suitable for analysis purposes. The preprocessing includes the following tasks [1,5,6,7]:

1. Data cleaning: This refers to the following tasks that ensure better data to be the input to the analysis process:
   a. Fill in missing values,
   b. Smooth noisy data,
   c. Identify or remove outliers,
   d. Resolve inconsistencies.
2. Data integration: This phase is used to integrate the heterogeneous data coming from different sources and different formats to be stored and located in one repository. This phase includes the following:
   a. Using multiple databases
   b. Data cubes
   c. Files.
3. Data transformation:
   a. Normalization
   b. Aggregation.

4. Data reduction: Data reduction represents one of the most important phase in data preprocessing since it aims to reduce data the volume as a step to improve the volume data prepared for the analysis process and hence improve the algorithm performance. The new data must produce the same or similar analytical results as the original.
5. Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Different preprocessing techniques were used to get clean data, these include [2, 3]:

1. Removing outliers, some of the data in datasets represent outliers and cannot be included in the analysis algorithms and techniques, so these data records were deleted from the set.
2. Filling missing data, some patient ages, jobs, and income were not mentioned in the tables, average and most commonly used values were used to substitute these missing values.
3. Data reduction using normalization and aggregation.

Reasons behind incompleteness, inconsistence and noisy data

1. Human errors
2. Machine and instrument errors
3. Heterogeneous data from different sources
4. Data transmission
5. Others

The diagram in figure (1) shows the preprocessing activities

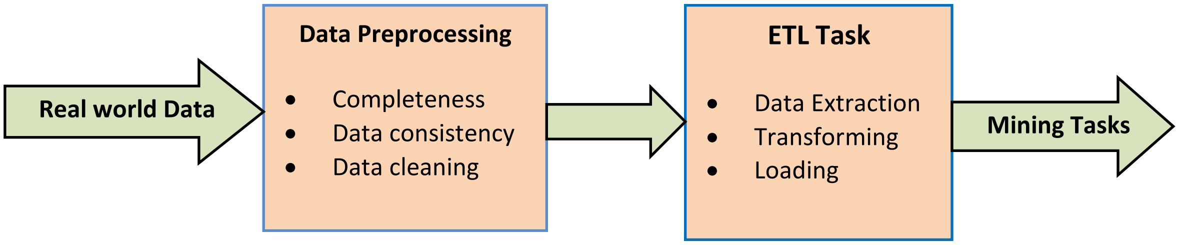


Figure (2): Data Preprocessing.

3. Data Collection

For the research paper, real data were collected from different Iraqi hospitals from many cities and locations in Iraq. Table (1) shows sample of the data collected were more than 20000 records represent the real data collected. Figure (3) shows the distribution of the original data under different attributes.

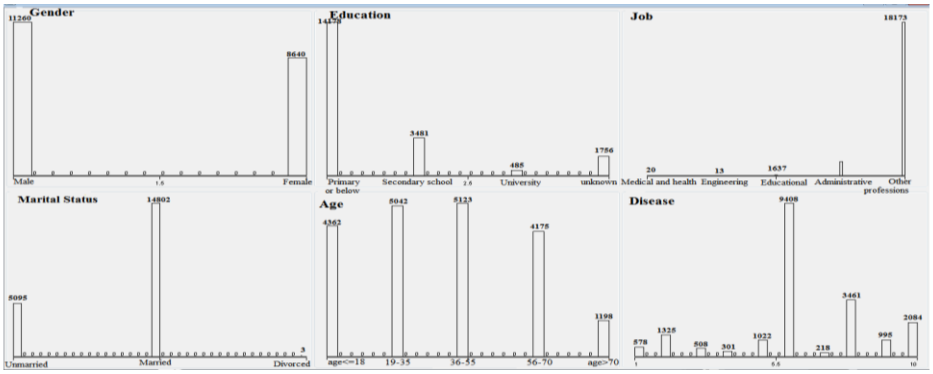


Figure (3) Distribution of the original data under different attributes

4. Missing values and their filling

Table (1) shows samples of the data set that contain some missing values in patient age attribute. It is required to replace these missing values in order to prepare the data for analysis process. Techniques in section 4.1 through 4.6 are used for this task [2, 3, 4].

4.1.Most common values in the data set

One of the most popular methods for replacing missing values in any data set is that by using the most common value in the whole data set. In our case for the patient data set the most common value as shown in table (2). [2, 3, 4].

4.2.Data set average to fill in missing values

Numeric attributes have more alternatives and algorithms for filling missing values in the data set, one of which is the overall average of the whole data set for the given attribute. However, in some cases this may lead to have results with high standard deviation from the original. For the data set used in this paper, the overall average for the age attribute shown in table(2). [2, 3, 4].

Table (1). Patient Data Set (Sample of 50 Records).

نو ع	المر ض	السكن	العم	الحالة	المه	التحصيل	الجنس	ر قم	التسل
0	H04	18		1	9	0	2	137	1
2	L87	18	20	1	13	2	2	204	2
2	O64	18	14	1	13	2	2	105	3
2	O64	18	2	1	9	0	1	107	4
2	E14	18	50	2	6	2	1	106	5
2	O36	18	7	1	8	4	2	221	6
2	O36	6	1	1	9	0	2	158	7
2	L87	18	16	1	13	2	2	97	8
2	O64	18	2	1	9	0	2	211	9
2	L02	18	27	2	6	5	1	215	10
3	O38	6		2	6	5	1	201	11
3	H61	18	45	2	6	2	1	241	12
0	S67	18	35	2	5	5	1	273	13
0	I20	18	60	2	13	2	2	25136	14
0	I20	18	82	2	10	1	2	25137	15
0	E14	18	65	2	13	2	2	25012	16
0	I21	18	52	2	6	2	1	25014	17
0	I70	18	60	2	13	2	2	12	18
0	I70	18	55	2	13	2	2	25080	19
0	I70	18	60	2	10	2	1	25102	20
0	I50	18	60	2	10	1	1	25158	21
0	I21	18	60	2	7	4	1	24645	22
0	I20	18	35	2	6	2	1	88	23
0	I50	18	65	2	10	1	2	9	24
3	N30	18	21	1	6	4	1	26	25
3	N41	18	60	2	10	5	1	25096	26
3	N35	18	15	1	6	4	1	24954	27
0	N20	13	70	2	10	5	1	27	28
3	N41	13	65	2	10	5	1	25075	29
3	K40	18	12	1	8	4	1	25097	30
3	N30	18	65	2	10	5	1	25117	31
3	N41	18	60	2	10	5	1	25065	32
3	N41	18	60	2	10	4	1	24951	33
3	N20	16	48	2	6	5	1	25115	34
3	N44	18	27	2	6	5	1	25094	35
3	N20	18	25	2	6	5	1	25093	36
3	C67	18	70	2	10	5	1	23922	37
3	C67	18	60	2	10	5	1	24883	38
0	N18	9	41	2	5	10	1	120	39
0	N18	18	42	2	13	2	2	25060	40
0	N18	18	66	2	10	2	1	142	41
0	N18	18	29	1	6	2	1	72	42
0	N18	18	42	2	13	2	2	24389	43
0	C67	18	59	2	10	5	1	25004	44
3	N44	18	25	2	6	4	1	1174	45
0	N35	18	54	2	5	5	1	25156	46
2	N21	18	65	2	10	1	2	180	47
2	N21	18	55	2	10	1	2	109	48
2	J03	18	17	1	8	6	1	101	49
0	N30	18	6	1	9	0	1	24865	50

Table (2). Results of Filling Ages Missing Values Using

Most Common Value and Data Set average

المعدل الكلي للاعمار		القيمة العمرية الاكثر		النسلسل
العمر	العمر	العمر	العمر	
39	1	60	1	1
39	6	60	6	2
39	9	60	9	3
39	10	60	10	4
39	14	60	14	5
39	19	60	19	6
39	20	60	20	7
39	25	60	25	8
39	29	60	29	9
39	30	60	30	10
39	35	60	35	11
39	39	60	39	12
39	40	60	40	13
39	43	60	43	14
39	48	60	48	15
39	50	60	50	16
39	55	60	55	17
39	59	60	59	18
39	60	60	60	19
39	64	60	64	20
39	68	60	68	21
39	70	60	70	22
39	72	60	72	23
39	77	60	77	24
39	85	60	85	25

4.2.Class most common values

To get better estimation for the missing values in the data set, more efficient algorithms are available, one of which is to classify the data set according to some specific attributes and in our cases, data set is classified according to the patient gender (male and Female) and the disease type. The most common values are then found for each class and these values are used to fill in the missing values in the objects under the same class. This method can be used for all data types of the attributes. See tables (3), (4),(5) and (6). [2, 3, 4].

4.3.Class average

For the numeric values attributes, the class average is an effective method for filling the missing values. Dividing the original table into smaller tables according to some predefined classes will lead to better estimation of the targeted missing values, as shown in tables (3), (4),(5) and (6). [2, 3, 4].

Table (3). Results of Filling Ages Missing Values Using Most Common Value and Data Set Average for Gender Class (Female)

معدل اعمار فئة الاناث		القيمة الاكثر شيوعا للاناث		التسلسل
العمر الجديد	العمر	العمر الجديد	العمر القديم	
40	6	60	6	1
40	8	60	8	2
40	9	60	9	3
40	11	60	11	4
40	14	60	14	5
40	17	60	17	6
40	20	60	20	7
40	25	60	25	8
40	26	60	26	9
40	30	60	30	10
40	35	60	35	11
40	38	60	38	12
40	40	60	40	13
40	43	60	43	14
40	45	60	45	15
40	50	60	50	16
40	55	60	55	17
40	59	60	59	18
40	60	60	60	19
40	63	60	63	20
40	65	60	65	21
40	70	60	70	22
40	71	60	71	23
40	80	60	80	24

4.4.Cluster most common values

When the attributes of the data to be classified are not defined prior to classification process, this is known as clustering. Putting patients into groups each group has some common characteristics (Similarities) between group objects and has dissimilarities with other groups is known as Clustering. The most common attribute value is used to replace the missing values. Non numeric attributes with missing values can be resolved using this algorithm. This part can be done as future work.

4.5.Cluster average

Attributes with numeric values and have missing values, cluster average can be used to fill in these missing values. Also this part can be done as future work.

5. Results and Conclusions

Filling in missing values in the data set prepared for analysis process is of great importance since it leads to clean data suitable for the analysis process under Data Mining Algorithms. Different data attributes have different techniques and algorithms for solving the missing values problem in real world data.

Missing values are very common in data analysis. In this research, more than 20000 records for the patients from different hospitals in different places in Iraq were collected. Different algorithms have been used to solve this problem, these include:

1. Most common values
2. Overall data set average
3. Class most common values
4. Class average
5. Clustering most common
6. Clustering average

According to the results in tables (II), (III), (IV), and (V), it is clear that classification algorithms are more suitable for replacing the missing values since the results are a much closer to the original data than any other algorithms.

Table (4). Results of Filling Ages Missing Values Using Most Common Value and Data Set Average for Gender Class (Male)

معدل اعمار فئة للذكور		القيمة الاكثر شيو عا		التسلسل No
العمر الحديد	العمر	العمر	العمر	
38	1	60	1	1
38	3	60	3	2
38	9	60	9	3
38	10	60	10	4
38	12	60	12	5
38	19	60	19	6
38	20	60	20	7
38	25	60	25	8
38	29	60	29	9
38	30	60	30	10
38	35	60	35	11
38	39	60	39	12
38	40	60	40	13
38	43	60	43	14
38	48	60	48	15
38	50	60	50	16
38	51	60	51	17
38	58	60	58	18
38	60	60	60	19
38	64	60	64	20
38	68	60	68	21
38	72	60	72	22
38	77	60	77	23
38	85	60	85	24

Table (5). Results of Filling Ages Missing Values Using Most Common Value and Data Set Average for (Gender + Disease) Class

معدل اعمار الفئة	العمر الاكثر شيوعا	العمرالقديم (القيم	الجنس	نوع المرض	التسلسل
51	50	39	M/ ذكر	E10	1
68	70	61	M/ ذكر	H25	2
68	70	55			3
62	65	38	M/ ذكر	I50	4
63	70	45	M/ ذكر	I68	5
63	70	85			6
63	70	68			7
63	70	64			8
25	20	12	M/ ذكر	K35	9
35	35	40	M/ ذكر	K62	10
45	45	32	M/ ذكر	N18	11
45	45	40			12
45	45	41			13
53	70	73	M/ ذكر	N30	14
22	30	50	M/ ذكر	N44	15
22	30	35			16
26	30	27	M/ ذكر	S07	17
26	30	2			18
26	30	31			19
32	35	45	F/انثى	B69	20
51	50	25	F/انثى	E14	21
66	60	80	F/انثى	H25	22
66	60	63			23
66	60	60			24
59	60	63	F/انثى	I20	25
64	60	60	F/انثى	I50	26
63	70	65	F/انثى	I68	27
63	70	70			28
63	70	50			29
14	7	31	F/انثى	J03	30
23	25	26	F/انثى	K36	31
41	35	59	F/انثى	K81	32
41	35	38			33
43	60	17	F/انثى	N18	34

Table (6). Results of Filling Ages Missing Values Using Most Common Value and Data Set Average for Job Class

المعدل للعمربعد اضافة	العمرالاكثرشيوعا بعداضافة	المهنة	العمرالقديم (القيم	التسلسل
41	50	6	39	1
69	70	10	61	2
64		5	55	3
80	80	11	38	4
49	55	5	45	5
69	70	10	85	6
			68	7
			64	8
15	12	8	12	9
37	40	6	40	10
38	45	6	32	11
			40	12
			41	13
66	70	10	73	14
31	30	5	50	15
			35	16
5	2	9	27	17
33	30	5	2	18
			31	19
35	35	13	45	20
42	50	13	25	21
66	60	10	80	22
			63	23
			60	24
64	60	10	63	25
67	60	10	60	26
68	70	10	65	27
			70	28
54	50	13	50	29
24	25	13	31	30
26	25	13	26	31
60	50	10	59	32
37	35	13	38	33
38	35	13	17	34

6. References

[1] M. Steinbach, P.-N.Tan and V. Kumar, Introduction to Data Mining, Addison-Wesley, 2006. ISBN: 0-321- 32136-7

[2] Jiawei Han and Micheline Kamber " Data Mining: Concepts and Techniques" 3rd Edition., Morgan Kaufmann, 2010.

[3] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.

[4]D. J. Hand, H.Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001.

[5] Deborah Osborne, MA, Susan Wernicke, MS, "Introduction to Crime Analysis: Basic Resources for Criminal Justice Practice, The Haworth Press, New York, London, Oxford, 2003.

[6] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed., 2005, ISBN 0-12-088407-0

[7]Tianyi Wu, Yuguo Chen and Jiawei Han, "Association Mining in Large Databases: A Re-Examination of Its Measures", in Proc. 2007 Int. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, Sept. 2007.