Estimating Null Values in Database Using CBR and Supervised Learning Classification

Khaled Nasser ElSayed
Computer Science Department, Umm Al-Qura University

***Abstract*—Database and database systems have been used widely in almost, all life activities. Sometimes missed data items are discovered as missed or null values in the database tables. The presented paper proposes a design for a supervised learning system to estimate missed values found in the university database. The values of estimated data items or data it items used in estimation are numeric and not computed. The system performs data classification based on Case-Based Reasoning (CBR) to estimate loosed marks of students. A data set is used in training the system under the supervision of an expert. After training the system to classify and estimate null values under expert supervision, it starts classification and estimation of null data by itself.**

Keywords—DataBase(DB);Data mining; Case-Based Reasoning (CBR); Classification;Null Values; Supervised Learning

I. INTRODUCTION

Database is a collection of related data, to represent some aspects of the real world, sometimes called the mini-world or the universe of discourse. It has become an essential component of everyday life in modern society. In the course of a day, most of us encounter several activities that involve some interaction with a database.

RDBSs are the mostly database systems used today. These system organize databases in many relations. Each relation has data about certain entity type or class and consists of rows. Each row represent a record of entity or object. The state of the whole database will correspond to the states of all its relations at a particular point of a time.

Data Mining is an essential process where intelligent methods are applied in order to extract data patterns. Data mining algorithms look for patterns in data. While most existing Data Mining approaches look for patterns in a single data table, relational Data Mining (RDM) approaches look for patterns that involve multiple tables (relations) from a relational database [1].

In recent years, the most common types of patterns and approaches considered in Data Mining have been extended to the relational case and RDM now encompasses relational association rule discovery and relational decision tree induction, among others. RDM approaches have been successfully applied to a number of problems in a variety of areas, most notably in the area of bioinformatics. This chapter provides a brief introduction to RDM [2].

Knowledge discovery in databases (KDD), also called data mining, has recently received wide attention from practitioners and researchers. There are several attractive application areas for KDD, and it seems that techniques from machine learning, statistics, and databases can be profitably combined to obtain useful methods and systems for KDD [3].

The KDD area should be largely guided by (successful) applications. Theoretical work in the area is needed. A KDD process in which the analyzer first produces lots of potentially interesting rules, subgroup descriptions, patterns, etc., and then interactively selects the truly interesting ones from these [4]

The presented system uses CBR classification in estimation null values in DB. The basic idea is locating a classified case (a student object) in the system Knowledge Base (KB) as the most close case to the student case row which has a null value. After that, the system could estimate that null value using three methods and their average.

The weight of each attribute is varied, to represent its effect in the total mark. The total mark at any moment is a resultant of the already registered marks in the fields of the table. At any time, the weight of each attribute it is computed as output of dividing the attribute value for a student by the resultant of maximum values of all registered attributes for that course.

Section II present some survey on related work. While, section III, outlines the structure of database record used by the system. Section IV, explores the system knowledge base. Section VI explains training the system and system classification experiment and results, while section VII discuss the conclusion and future work.

II. RELATED WORK

A lot of research effort have been done in estimating null values in DB. The pioneers, Chen, in this area used a new method to estimate null values in relational database in [5]. They improved their method by creating fuzzy rule base in [6] and used genetic algorithms for generated weighted fuzzy rules in [7]. Then, they applied the automatic clustering algorithm for clustering the tuples in the relational database in [8]. Then, they presented a new method for estimate null values in relational database systems having negative dependency relationships between attributes in [9], where the "Benz secondhand car database" is used for the experiment.

Wang, C.H. Cheng, and W.T. Chang [10] utilized stepwise regression to select the important attributes from the database and a partitional approach to build the datacategory. They apply the clustering method to cluster output data. Also, Chen and Hsaio [11] and Cheng and Lin [12] utilized clustering algorithms to cluster data, and calculate coefficient values

between different attributes by generating minimum average error.

Jain and Suryawanshi [13] proposed an efficient approach for handling null values in web log. They used Tabu search–KNN classifier perform featureselection of K-NN rules.Also, C.H Cheng, J.R. Chang, and L.Y. Wei in [14] used adaptive learning techniques, based on clustering, toresolve the issue of null values in relational database systems. This study uses clustering algorithms to group data and calculates the degree of influence between independentattributes (variables) and the dependent attribute through an adaptive learning method.

Lee and Wang in [15] proposed a modular method for trying to processhigh-reliability relational database estimation, and thestructure of the proposed method can be composed of threephases, comprising partition determination, automatic fuzzysystem generation, and relational database estimation. While, Mridha and Banik used Noble evolutionary algorithm to generating weighted fuzzy rules to estimate null values [16].

Sadiq, S.A. Chawishly, and N.J. Sulaka in [17] proposed a hybrid approach for solving null values problem, it hybridize rough set theory with ID3 (Iterative Dichotomiser 3) decision tree induction algorithm. The proposed approach is a supervised learning model. Large set of complete data called learning data is used to find the decision rule sets that then have been used in solving the incomplete data. Then, the intelligent swarm algorithm is used for feature selection which represents bees algorithm as heuristic search algorithm combined with rough set theory as evaluation function[18].

III. Database Application

The proposed approach is tested in relational data base (DB) of university students. This database consists of many relations. Each relation is concerned of certain records of entity set. The target table is the STUDY table, shown in table 1, which concern of the remarks and grades of students in the registered courses.

Sometimes there missed or null values in a column of certain records in databases. As Example some remarks data of some student exams are missed. These null values might result from missing some exam grades or from non-entering mistakes.

As example, the estimation of null values is applied for a course has the assessments: two quiz (q1 and q2), five home works, a project, midterm exam, final exam. But it is possible to add or remove some assessment(s) to/from the proposed list of assessment(s). The STUDY table has those attributes, as shown in Table 1, which shows some records of student remarks.

The experiment is applied over marks data of the course, “Compilers Construction" in Computer Science department. Table 2, presents the universe of discourse of the attributes Home works, quizzes, MidTerm, Project, and Final Exam.

The attributes (column) of any database entry (SQL table) that have null values, are classified into four types, according to the reason and type of missing values or the ability of estimating the null values.

1) ***Type$_1$*** *is NullColumn,* where any column, like MedTerm as example, may have all of its values are null. This means that the column values are not inserted or computed yet.

2) ***Type$_2$*** *is NotEstimated*, where the attribute null values can't be estimated by any system. As example, the attributes: Student_num(St#), Name, Address, or Cours_Code.

3) ***Type$_3$*** *is Derived or Computed*, like Total. The attribute null value can be computed or imaged from another attribute(s).

The action in the first three types is running the program that computes or acquires those null data, or fill them by user.

4) ***Type$_4$*** *is Can-Estimated*, where a value of an attribute in certain row(student record) is missed. This null value can be estimated by the system.

TABLE I. Study Table With Actual Values of Homeworks, Quizzes, Midterm, Project, and Final Exams.

St#	Q1 /5	Q2 /5	H1/ 2	H2 /2	H3 /2	MTerm /20	H4 /2	H5 /2	Project /20	Final /40	Total /100
1	3	4	2	1	1	18	1	1	15	36	79
2	2	3	2	2	2	18	2	1	15	38	83
3	1.5	2.5	2	2	2	1	2	1	15	39	66.5
4	5	4	2	1	1	11	1	1	15	2	38
5	5	5	2	1	2	17	1	2	17	24	71
6	3	2	2	1	2	17	2	1	17	26	70
7	4	3	0	2	1	18	2	1	17	2	46
8	2	3	0	2	2	18	1	2	17	28	73
9	1	1	0	2	2	18	2	2	12	0	39
10	3	3	1	1	1	19	2	2	12	0	41
11	5	4	1	2	2	19	1	2	12	35	78
12	3	5	2	1	1	18	1	1	12	37	78
13	4	5	2	2	1	7	1	2	0	8	28
14	4	5	2	1	2	17	2	2	0	36	67
15	4	3	2	2	1	16	2	1	0	33	60
16	4	3	2	1	2	16	2	2	0	32	60

TABLE II. Universe of Discourse For Home Works, Quizzes, Midterm, Project, and Final Exams

Attribute Name (Assessment)	Minimum Value	Maximum Value
Home Works (H_1,H_2,H_3,H_4,H_5)	0	2
Quizzes (Q_1,Q_2)	0	5
MidTerm Exam	0	20
Project	0	20
Final Exam	0	40

The proposed method estimates null values in all column of type 4, based on the values of the known marks in the database. Thus, the known and estimated values are numeric values. Then system computes the total remark.

IV. Knowledge Representation

The system should acquires basic knowledge needed to build its KB, shown in Fig. 1. This process trains the system to

learn classification and estimation under the supervision of the expert. Fig. 2 demonstrates the algorithm for this process.

Each student object is scanned to be classified is stored as a case. Each case is described by its attributes of certain row in Table 1. Values of these attributes will be used in classification (category) of student objects. The category gives impression about the level of student objects related to it. It refers to the range within which their total resultant of registered attributes grades divided by the total of maximum marks of those attributes, in certain course. It has actual categories like: APLUS, A, BPLUS,B,……..,FAIL, LOWFAIL.

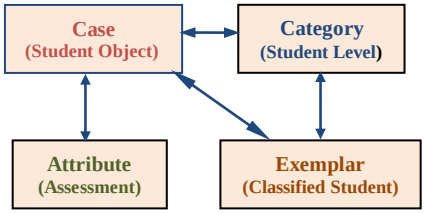

Fig. 1. System Knowledge Base

Each classified student object is related to a category. It is known as an exemplar of that category. It is represented as a combination of values of assessments attributes. There is no restriction on number or names of categories and exemplars.

Student object, Student Level, Attribute, and exemplar are represented as C++ classes. These classes and their relationships represent the knowledge base of the proposed system.

V. Training and Supervised Learning Process

At running the system for the first time, it reads the student objects (rows of Table) and checks there attributes for null values. Then, it gives report of null value types according to the preliminary classification given in section III. Also, it specifies the rows which has null values to be estimated later as described in section VI.

For each attribute a scale of possible values is determined. The combination of all possible attribute values defines all possible marks states within this description. The task is to classify each student object's state.

When a student objects (cases) are scanned by the system - for the fist time - to classify, it can do nothing. It has no categories and no classified exemplars to match. It'll ask for help from the expert to classify and clarify reason for that classification. First distinct cases will be classified by the expert and added to KB as exemplars. Those exemplar are related to new created categories.

At reading a new row of student object (unclassified case), the system will start classification process to specify a category from KB categories, based on values of its attributes. If the category is not in the KB yet, it will ask the expert to create new one, and name it. Categories names are listed section IV. Within each category, there will be many exemplar, each has its level. This level should be within the space of the category.

Exemplar level =*sum of actual values of all encountered attributes*/*sum of maximum values of all encountered attributes*.

For a new case, the system looks up for a similar exemplar to it. If it finds a category, it consult its suggestions to the expert. If the expert accepts, the new case is related to the category and a new exemplar is created if expert want. While, if the expert refuses that classification, or the system fails to find a category, it asks the expert to explain why? And classify himself and give reasons. Then a new category and an exemplar (new case) related to that category are created.

The expert may classify the new case to an existing category, or even a new one. The Algorithm of system training and classification is shown in Fig. 2. Supervised learning will continue in the estimation process, as seen in the next section.

Discard column of $type_1$, $type_2$, $type_3$, and $type_4$.
For each row of complete registered data Do

1. **Read actual and maximum values for each attribute of encountered column.**
2. **Compute the Total of Actual Values (TAV).**
3. **Compute the Total of Maximum Values (TMV).**
4. **Compute Case Level (CL) = TAV/MTV.**
5. **If there is no Category, its scope include CL, Consult expert to get name and scope of new Category, and Create it.**
6. **Exemplar level (EL) = CL.**
7. **Add new Exemplar related to the specified category.**
8. **Ask Expert for Classification reason and Learn.**

Fig. 2. Algorithm of System Training and Classification.

VI. Estimating Null Values

A. System Classification of Student Object

Mainly, the use of CBR classification is for locating the most close case (exemplar) to the student case which has null data. Student object of null value is the object has to be classified and assigned to a certain class (category). It is constructed from its marks of assessments (attributes) and their weight. It is clear that an expert Instructor uses that knowledge to characterize a marks condition. Assuming that in order to make preliminary conclusions the expert uses a finite number of marks of assessments.

Each attribute has a weight, based on the its space of minimum and maximum marl value that can be assigned to it, compared with the total of values all registered attributes for students.

Some attribute may be not available at certain moment for a course. There may an assessment canceled, or not held yet. So, the N/A attributes should excluded from the list of attributes that describe a student objects (cases) for a while, until be included in the DB. This happens as done with column of $type_1$, $type_2$, and $type_3$.

When a new case (student object), with null value in one of its attributes, is found out, the system start its classification process. It looks up for a category for that case and discovers the most similar exemplar (classification) to that case. If it fails, it asks for expert classification. While, if it successes, it starts estimating of the null value for the current classified student (case). The Algorithm of system classification and estimation is presented in Fig. 3.

After classifying the student object to be related to certain category, the system retrieves the exemplars related to the same category. It might use one of four methods to estimate null values.

1. **Read actual and maximum values for the student row where it has a null value.**
2. **Compute the Total of Actual Values (TAV).**
3. **Compute the Total of Maximum Values (TMV).**
4. **Compute *Case Level (CL) = TAV/MTV*.**
5. **If there is no Category, its scope include CL, Consult expert to get name and scope of new Category, and Create it.**
   Else locate category and all exemplars related to it
6. **Estimate Value for null Value using three estimation methods (Est_1, Est_2, and Est_3).**
7. **Compute the *EstAvg= (Est_1, Est_2, and Est_3)/3*.**
8. **Consult Estimation Values (Est_1, Est_2, Est_3, and EstAvg) to the Expert, to select one.**
9. **Ask Expert for selection reason, and Learn**.

Fig. 3. Algorithm for System Training and Classification.

Value of null attribute A in the current student record is estimated as any of the following methods:

1) The opposite value of the same attribute A in the most similar exemplar.

2) The average of all opposite values for attribute A in all exemplars related to the classification category.

*3) The average of level of all exemplars related to the classification category * maximum value of that attribute (out of marks).*

4) The average of the results from 1,2 and 3.

Then, the system offers its estimated values to the expert, to get his selection and guidance. The expert should choose one of them or refuse all. For all chooses, the system ask the expert for reasons of his decision.

Most of times, the expert reason was that the selected method is suitable for the nature, weight, and difficulty of each assessment (attribute). Next times, the system will use this knowledge to choose the method itself. Comparing results of estimating for assumed null values attribute will explain next. Finally, the system calculates the average of all methods. As seen next.

B. Experiment Results

Assume that there are n records (R_1, R2,….,R_n) in the STUDY table of the database, where the value of the attribute "MidTerm" of the record R_i is "R_i.MidTerm", as example. Also, assume that the estimated values of Ri.Midterm are ER_i.MidTerm (method1, method2, method3, method4). To estimate the value of the missed MidTerm value, those four values are estimated according to the four methods listed above.

Referring to the table STUDY showed in table 1, and assuming that there is null value in a certain records. five assumptions will be tested, while Table 3, collects results of the following experiment to estimate null values in an attribute of five columns of different records:

1) The record of 15th student record has null value in the column of MidTerm, while other attributes are given their values.

2) The record of 5th student record has null value in the column of homework H1, while other attributes are given their values.

3) The record of 8th student record has null value in the column of Final Exam, while other attributes are given their values.

4) The record of 10th student record has null value in the column of quiz Q1, while other attributes are given their values.

5) The record of 12th student record has null value in the column of Project, while other attributes are given their values.

TABLE III. RESULTS OF THE EXPERIMENTS

Experiment	Method1 value	Method2 Value	Method3 value	Average Value	Actual Value
ER_{15}.MedTerm	16	16	12	14.66	16
ER_5.H1	2	1	1.41	1.47	2
ER_8.FinalExam	24	25	18.5	22.53	28
ER_{10}.Q1	4	3.25	1.88	3.04	3
ER_{12}.Project	12	13.5	13	12.83	12

As seen in table 3, there is no method is preferred to applied for all attributes. While, the average of all estimation process, is somehow reasonable and applicable. Also, it is noticed that if the number of rows increases, the precision of the estimation will increase also.

VII. CONCLUSIONS

This paper presented a supervised learning system for estimating null values found in the database. The system performs data classification based on CBR-based classification to estimate missed marks of students. A moderate data set is used in training the system under the supervision of an expert, then the system start classification of objects that have null values using four methods. It is found that the average of the estimated values is more reasonable and applicable. In future, improvements will be applied to increase the precession of estimated values. Bigger training data set will be used in training the system to improve precision.

Also, the task of estimation will be enlarged to enable the system to estimate a multiple null values not only one null value in the in the same record.

REFERENCES

[1] J. Han & M. Kamber “Data Mining Concepts and Techniques”, 2nd edition, The Morgan Kaufmann Series in Data Management Systems Series Editor: Jim Gray, Microsoft Research Data Mining, ElServierInc, 2006.

[2] S. Džeroski “Data Mining and Knowledge Discovery Handbook”, Part 6, 887-911, Springer, 2010.

[3] H. Mannila& H. Toivonen “Levelwise Search and Borders of Theories in Knowledge Discovery”, Data Mining and Knowledge Discovery 1, 241–258, Kluwer Academic Publishers, Manufactured in The Netherlands, 1997.

[4] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, R. (Eds.), “Advances in Knowledge Discovery and Data Mining”, Menlo Park, CA: AAAI Press, 1996.

[5] S.M. Chen and H.H. Chen, "Estimating Null Values in the Distributed Relational Databases Environment", Cybern, Syst., Vol. 31, No. 8, pp. 851-871, 2000.

[6] S.M. Chen, S.H. Lee, and C.H. Lee, "A New Method for Generating Fuzzy Rule from Numerical Data for Handling Classification Problems", App. Art. Intell., Vol. 15, No. 7, pp. 645-664, 2001.

[7] S.M. Chen and C.M. Huang, "Generating Weighted Fuzzy Rules from Relational Database Systems for Estimating Null Values using Genetic Algorithms", IEEE Transactions on Fuzzy Systems, Vol. 11, No. 4, pp. 495-506, August 2003.

[8] S. M. Chen and C. M. Huang, “A new approach to generate weighted fuzzy rules using genetic algorithms for estimating null values,” Expert Systems with Applications, vol. 35, no. 3, pp. 905-917, October 2008.

[9] S.M Chen and S.T. Chang, "Estimating Null Values in Relational Database Systems Having Negative Dependency Relationships Between Attributes", Journal Cybernetics and Systems, , Vol. 40, No. 2, pp. 146-159, February 2009.

[10] J.W. Wang, C.H. Cheng, and W.T. Chang, "Partitional Approach for Estimating Null Value in Relational Database", Springer-Verlag AI 2005, LNAI 3809, pp. 1213-1216, 2005.

[11] S.M. Chen and H.R. Hsiao, "A new method to estimate null values in relational database systems based on automatic clustering techniques", Elsevier Inc., Information Sciences 169, pp. 47-69, 2005.

[12] C.H. Cheng and T.C. Lin, "Improving Relational Database Quality Based on Adaptive Learning Method for Estimating Null Value", ICICIC, 2007, Innovative Computing, Information and Control, International Conference on, Innovative Computing, Information and Control, International Conference on 2007, pp. 81-89.

[13] Y. K. Jain and V. Suryawanshi, "A New Approach for Handling Null values in Web Log Using KNN and Tabu Search KNN", International Journal of Data mining &Knowledge Management Process (IJDKP), Vol. 1, No. 5, pp.9-19, September 2011.

[14] C.H Cheng, J.R. Chang, and L.Y. Wei, "ADAPTIVE-CLUSTERING BASED METHOD TO ESTIMATE NULL VALUES IN RELATIONAL DATABASES", International Journal of Innovative Computing, Information and Control(ICIC), Vol. 7, No. 1, pp. 223-235, January 2011.

[15] S.J. Lee and H.S. Wang, "A Dynamic Modular Method for Estimating Null Values in Relational Database Systems", International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), Vol. 1, pp. 249-257, 2009.

[16] M.F. Mridha and M. Banik, "Performances of Estimating Null Values using Noble Evolutionary Algorithm (NEAs) by generating Weighted Fuzzy Rules", International Journal of Computer Applications, Vol. 11, No. 9, pp. 30-35, December 2010.

[17] A.T. Sadiq, S.A. Chawishly, and N.J. Sulaka, "Intelligent Methods to Solve Null Values Problem in Databases ", Journal of Advanced Computer Science and Technology Research, Vol. 2, No. 2, pp. 91-103, June 2012.

[18] A.T. Sadiq, M.G. Duaimi, and S. A. Shaker, "Data Missing Solution Using Rough Set Theory and Swarm Intelligence", International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 3, PP. 1-16,2013.

AUTHOR PROFILE

The Author is Dr. Eng. Khaled N. ElSayed. He was born in Cairo, Egypt 9 Oct. 1963. He have got his PhD of computers and systems fromFaculty of Engineering, Ain Shams University, Cairo, Egypt, 1996.

He has worked as an associate professor of computer science, in Umm-AlQura Uni. in Makkah, Saudi Arabia since 2006. Artificial Intelligence is his major. His interest research is Distant Education, E-Learning, and Agent.

Dr. Khaled N. ElSayed translated the 4th edition of “Fundamentals of Database Systems”, RamezElmasei and Shamkant B. Navathe, Addison Wesley, fourth edition, 2004, published by King Saud University, Riyadh, Saudi Arabia, 2009. He is also the author several books in programming in C & C++, Data structures in C& C++, Computer and Society, Database Design and Artificial Intelligence.