International Journal of Innovative Computing, Information and Control
Volume **7**, Number **1**, January **2011**

ICIC International ©2011 ISSN 1349-4198
pp. **223–235**

ADAPTIVE-CLUSTERING BASED METHOD TO ESTIMATE NULL VALUES IN RELATIONAL DATABASES

Ching-Hsue Cheng[1], Jing-Rong Chang[2,*] and Liang-Ying Wei[3]

[1]Department of Information Management
National Yunlin University of Science and Technology
No. 123, Sec. 3, University Rd., Touliu, Yunlin 640, Taiwan
chcheng@yuntech.edu.tw

[2]Department of Information Management
Chaoyang University of Technology
No. 168, Jifong East Rd., Wufong Township, Taichung County 41349, Taiwan
*Corresponding author: chrischang@cyut.edu.tw

[3]Department of Information Management
Yuanpei University
No. 306, Yuanpei Street, Hsin Chu 30015, Taiwan
lywei@mail.ypu.edu.tw

Received August 2009; revised May 2010

Abstract. *Data preprocessing is an essential step of knowledge discovery. Data preprocessing comprises data cleaning, data integration, data transformation, data reduction and data discretization. Estimating null values is a task of data cleaning. Null values in a database are significant sources of poor data quality. Therefore, the appropriate handling of null values is an important task of data preprocessing in relational databases. We propose a new method that uses adaptive learning techniques, based on clustering, to resolve the issue of null values in relational database systems. This study uses clustering algorithms to group data and calculates the degree of influence between independent attributes (variables) and the dependent attribute through an adaptive learning method (the best adaptive parameter can be obtained by the minimum average error rate). Three databases (a human resource database, Waugh's database and a government salary study database) were selected as the experimental data to compare the mean absolute error rate (MAER) of the proposed algorithm with the other methods. The results demonstrate that the proposed method outperforms other methods.*

Keywords: Relational database systems, Null value, Degree of influential, K-means, Adaptive learning

1. **Introduction.** In the information age, relational database systems play an important role in the business processes of enterprises. Yet, null values exist in the systems, which lead to system failure or create difficulties in completing important tasks, such as financial analysis and identification of target markets. Further, inaccurate, inconsistent and null-value attributes might be present in the database, hindering a researcher's ability to discover useful knowledge. An effective data quality strategy can help researchers locate information in a database, make the correct decision and reduce costly operational inefficiencies. In addition, the handling of null values is a major task in data preprocessing during data mining [1-6].

A database system will not operate properly if null values of attributes (incomplete datasets) exist in the system [7]. According to Han [8], null values arise due to one of five possibilities: (a) the data were not captured, due to faulty equipment; (b) inconsistencies