A Non-Parametric Discretization Based Imputation Algorithm for Continuous Attributes with Missing Data Values

G Madhu\textsuperscript{a}, T V Rajinikanth\textsuperscript{b}, A Govardhan\textsuperscript{c},

\textsuperscript{a}Department of Information Technology, VNR VJIET, Hyderabad–500090 India, Contact: madhu_g@vnrvjiet.in
\textsuperscript{b}Department of Computer Science and Engineering, SNIST Hyderabad–501301, India.
\textsuperscript{c}School of Information Technology, JNT University, Hyderabad–500085, India

Many real world data sets predominantly consist of numeric attributes with missing datasets. Supervised learning tasks involve numeric or continuous attributes. Consequently, appropriate handling of continuous attributes with missing data values is an important issue in the data mining process and machine learning perspective. Recently, many of the researchers have been proposed several supervised learning algorithms to handle only nominal attributes, continuous attributes or both but not numeric or continuous attributes with missing data values. To handle, continuous values with many discretization algorithms have been proposed in the literature, but not on numeric or continuous attribute with missing data values. In this paper, we propose a new non-parametric discretization based imputation algorithm for continuous attributes with missing data values using a popular statistical technique z-score with an index measure to impute the missing data values for numeric or continuous attributes. The experimental results show the proposed non-parametric discretization based imputation algorithm significantly enhances the efficiency in terms of accuracy and to minimize the classifier confusion of missing data values of continuous attributes in machine learning classifiers.

Keywords : Classification, Continuous Attributes, Discretization, Imputation, Missing Data Values.

1. INTRODUCTION

Real-world datasets frequently occur one common problem, incomplete or missing data values \cite{1, 2, 3}, these datasets are usually collected from different type of sources, like medical databases, astronomical, sensor networks, information repositories and others. Consequently, these datasets predominantly consist of continuous attributes also known as quantitative or numerical attributes (real or integer) and nominal attributes. The continuous or numeric attributes need to be transformed into discrete data values for data mining and machine learning task. Here, the transformation procedure is known as discretization, discretization play a vital role in machine learning and data mining algorithms for the past decade \cite{4}. Discretization was first discussed for qualitative data in classification learning algorithms \cite{5} \cite{6} \cite{7}.

In the literature many authors have proposed different discretization techniques, which applied to other datasets but, not on continuous missing data values. On the other hand, to deal with the missing value problem, many authors have proposed into two types of methods, i.e., ignored (deleted) or imputed (filling in) with suitable values \cite{8} \cite{9}. In \cite{10} \cite{11} discussed regression imputation, Hot-Deck Imputation \cite{12}, Imputation with K-Nearest Neighbor algorithm \cite{13}, K-means Clustering Imputation algorithm \cite{14}, Imputation with Fuzzy K-Means Clustering \cite{15}, Weighted imputation with K-Nearest Neighbor \cite{16}, Support Vector
tion with index measure imputation algorithm. The proposed algorithm to transform the continuous values into discrete one’s and imputes the missing attribute with new index measure algorithm. The proposed non-parametric discretization imputation algorithm has outperformed the state-of-the-art imputation methodologies on mixed missing attribute datasets considered in our experiments. We used Wilcoxon signed ranks test on this algorithm to test the performance in terms of classification accuracy and computational complexity on continuous or real-valued attributes with missing data values instead of traditional imputation algorithms. Finally, we conclude that our new non-parametric discretization based imputation algorithm is superior to other traditional imputation algorithms in terms of accuracy as well as computational complexity.

Acknowledgments. The authors would like to thank Prof. V Sree Hari Rao and Dr. M Naresh Kumar for their valuable suggestions.

REFERENCES
6. Kerber R and Chimerge. Discretization for Nu-
A Non-Parametric Discretization Based Imputation Algorithm


Dr. A Govardhan did his BE in Computer Science and Engineering from Osmania University College of Engineering, Hyderabad, M.Tech from Jawaharlal Nehru University, Delhi and Ph.D from Jawaharlal Nehru Technological University, Hyderabad. He is presently working as Director, SIT, JNTU Hyderabad, A.P, INDIA. He has 63 research publications at International/National Journals and Conferences. He is also a reviewer of research papers of various Journals. His areas of interest include Databases, Data Warehousing and Mining, Information Retrieval, Computer Networks, Image Processing and Object Oriented Technologies.