International Journal of Information Processing, 8(1), 64-72, 2014
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

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

G Madhu[a], T V Rajinikanth[b], A Govardhan[c],

[a]Department of Information Technology, VNR VJIET, Hyderabad–500090 India,
Contact: madhu_g@vnrvjiet.in

[b]Department of Computer Science and Engineering, SNIST Hyderabad–501301, India.

[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 [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 [4]. Discretization was first discussed for qualitative data in classification learning algorithms [5] [6] [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 [8][9]. In [10][11] discussed regression imputation, Hot-Deck Imputation [12], Imputation with K-Nearest Neighbor algorithm [13], K-means Clustering Imputation algorithm [14], Imputation with Fuzzy K-Means Clustering [15], Weighted imputation with K-Nearest Neighbor [16], Support Vector

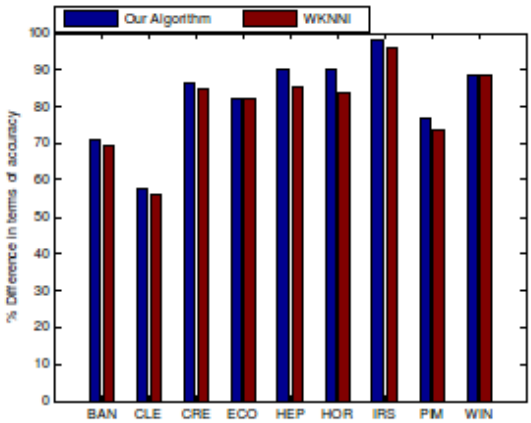

Figure 3. Our Algorithm Vs. WKNNI.

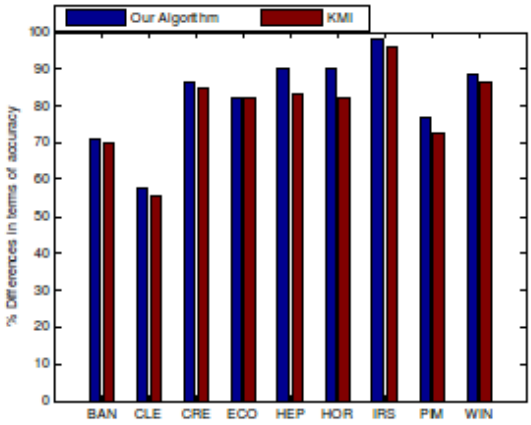

Figure 4. Our Algorithm Vs. KMI.

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

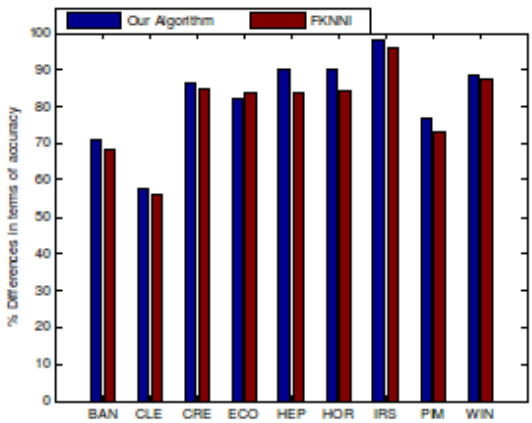

Figure 5. Our Algorithm Vs. FKNNI.

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

1. Allison P D. Missing Data, Sage University Papers Series on Quantitative Applications in the Social Sciences, *Thousand Oaks, California, USA,* 2001.
2. Duda R O. Pattern Classification, *Wiley-Interscience, New York,* 2000.
3. Little R J A and Rubin D B. Statistical Analysis with Missing Data, *Second Edition, Wiley NJ, USA,* 2002.
4. Cheng-Jung Tsai, Chien-I. Lee and Wei-Pang Yang. A Discretization Algorithm Based on Class-Attribute Contingency Coefficient, *Information Sciences: an International Journal,* 178:714–731, 2008.
5. Dougherty J, Kohavi R and Sahami M. Supervised and Unsupervised Discretization of Continuous Features. *In Proceedings of the 12th International Conference on Machine Learning,* pages 194–202, 1995.
6. Kerber R and Chimerge. Discretization for Nu-

meric Attributes, *In National Conference on Artificial Intelligence, AAAI Press,* pages 123–128, 1992.
7. Kohavi R and Sahami M. Error-based and Entropy-based Discretization of Continuous Features, *In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining,* pages 114–119, 1996.
8. Farhangfar A. A Novel Framework for Imputation of Missing Values in Databases, *IEEE Transaction on Systems, Man Cybernetics Part A: System Humans,* 37(5):692–709, 2007.
9. Pedro J Garc, a-Laencina. Pattern Classification with Missing Data: A Review, *Neural Computing and Applications,* 19(2):263–282, 2010.
10. Frane J W. Some Simple Procedures for Handling Missing Data in Multivariate Analysis, *Psychometrika,* 41:409–415, 1976.
11. Cohen J, Cohen E. Applied Multiple Regression/Correlational Analysis for the Behavioral Sciences, *(2nd ed.), Hillsdale, Erlbaum, NJ,* 1983.
12. J N K Rao, J Shao. Jackknife Variance Estimation with Survey Data under Hot Deck Imputation, *Biometrika,* 79(4):811–822, 1992.
13. G E A P A Batista, M C Monard. An Analysis of Four Missing Data Treatment Methods for Supervised Learning, *Applied Artificial Intelligence,* 17:519–533, 2003.
14. Li D, Deogun J, SpauldingW and Shuart B. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method, *In Proceedings of the 4th International Conference of Rough Sets and Current Trends in Computing (RSCTC),* pages 573–579, 2004.
15. Acuna E and Rodriguez C. The Treatment of Missing Values and its Effect in the Classifier Accuracy, Classification, Clustering and Data Mining Applications, *Springer, Berlin,* pages 639–648, 2004.
16. Troyanskaya O, Cantor M and Sherlock G. Missing Value Estimation Methods for DNA Microarrays, Bioinformatics, 17(6):520–525, 2001.
17. H Feng, C Guoshun, Y Cheng, B Yang and Y Chen. A SVM Regression Based Approach to Filling in Missing Values, *in Proceedings of KES*, 3:581–587, 2005.
18. Luengo J, Garca S and Herrera F. A Study on the Use of Imputation Methods for Experimentation with Radial Basis Function Network Classifiers Handling Missing Attribute Values: The Good Synergy Between RBFNs and Event Covering Method, *Neural Nets,* 23(3):406–418, 2010.
19. Richard Butterworth. A Greedy Algorithm for Supervised Discretization, *Journal of Biomedical Informatics,* 37:285–292, 2004.
20. Fayyad U M and Irani K B. Multi-Interval Discretization of Continuous-Values Attributes for Classification Learning, *in Proceedings of 13th International Joint Conference on Artificial Intelligence,* pages 1022–1027, 1993.
21. Sturges H A. The Choice of a Class Interval, *Journal of American Statistical Association,* pages 65–66, 1926.
22. I Witten, E Frank. Data Mining: Practical Machine Learning Tools and Techniques, *Morgan Kaufmann, San Francisco,* 2005.
23. J Alcala-Fdez. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework, *Journal of Multiple-Valued Logic and Soft Computing,* 17(2-3):255–287, 2011. [Online]. Available: http:// http://www.keel.es/

G Madhu is currently working as Associate Professor, Dept of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad-90. He obtained his Bachelor of Degree from Kakatiya University. He received his Masters degree in Computer Science and Engineering from Jawaharlal Nehru Technolgoical University Hyderabad. He is a professional member of ISTE, ISRS.

Dr. T V Rajinikanth has obtained his Ph.D Degree in CSE branch from Osmania university, Hyderabad in July, 2008 and MTech.(CSE) degree from Osmania University, Hyderabad in January, 2001. His specialization area in research is "Spatial Data- Mining". He obtained his PGDCS degree from HCU, Hyderabad in 1996. He received his MSc. (Applied mathematics) degree in the year 1989 from S V University, Tirupati as University Ranker. He is currently working as professor in CSE, at SNIST, Hyderabad. His current research area interests include

Image processing, Data Warehousing and Mining, Spatial Data Mining, Web Mining, Text Mining and Robotic Area *etc.*.

Dr. A Govardhan did his BE in Computer Science and Engineering from Osmania University College of Engineering, Hyderabad, MTech from Jawaharlal Nehru University, Delhi and Ph.D from Jawaharlal Nehru Technological University, 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.