Abstract

Knowledge Discovery in Dataset (KDD) plays a vital role in information analysis and retrieval based applications. Quality of data is the most indispensable component of KDD. The factor which affects the quality of datasets is presence of missing values. The data collected from the real world often contains serious data quality troubles such as incomplete, redundant, inconsistent, and/or noisy data. Handling missing values should be cautiously considered, or else prejudice might be introduced into the knowledge induced. The current work investigates three different treatments for dealing with missing values in United States Congressional Voting Records Database. All the machine learning methods were employed in one of the leading open-source data mining applications. This proposed study centers on the performance Evaluation of several classification models induced from data after applying three different methods to treat missing values. Results show that by boosting the k-nearest neighbor for imputation bids significant enhancement over traditional techniques (case/pairwise deletion and Replace missing value using mean ).

References

An Enhanced Approach for Treating Missing Value using Boosted K-NN

Computer Science (IWCECS'09).
- Data mining handling missing value http://www.developerzenom/2009/08/14/data-mining-handling-missing-values-the-database/
- Bernadette Bouchon-Meunier, Marcin Detyeniecki, Marie-Jeanne Lesot, Christophe Marsala, and Maria Rifqi. : Real-World Fuzzy Logic Applications in Data Mining and Information Retrieval: Page(s) 219-247
- Bikash Mukhopadhyay, Sripati Mukhopadhyay, : Data Mining Techniques for Information Retrieval 2nd International CALIBER-2004, New Delhi, 11-13 February, 2004
- Jaana Kekäläinen & Kalervo Järvelin, "evaluating information retrieval system under the challenges of interaction of multidimensional dynamicrelevance", Published in: Harry Bruce, Raya Fidel, Peter Ingwersen, and Pertti Vakkari (Eds.) Proceed-ings of the 4th CoLIS Conference. Greenwood Village, CO: Libraries Unlimited, Page(s). 253-270.
- James Allan (editor), Jay Aslam, Nicholas Belkin, Chris Buckley, Jamie Callan, Bruce Croft (editor), Sue Dumais, Norbert Fuhr. : Challenges in Information Retrieval and Language Modeling. : Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002
- Bayes Theorem, "http://www.cuttheknot.org/Probability/BayesTheorem.shtml"
- chao-ying joanne peng, kuk lida lee, gary m. Ingersoll. : An Introduction to Logistic Regression Analysis and Reporting: The Journal of Educational Research: Page(s) 3-14
- Tejaswini Abhijit Hilage, R. V. Kulkarni. : Review of literature on data mining. IJRRAS 10 (1) January 2012 Page(s) 107-114
- "precision and recall", http://www.bainsight.com/blog-archive/Pages/3-7-2011-1.aspx
- Precision and accuracy, "http://www.worsleyschool.net/science/files/precision/andaccuracy.html".
- Virpi Lyytikäinen, Pasi Tiitinen, Airi Salminen. : Challenges for European legal information retrieval Published in F. Galindo & G. Quirchmayer (Eds.), Proceedings of the IFIP 8. 5 Working. Page(s) 1-16
Page{1721-1726
- Raghda Fouad, Mohamed Hashem, Nagwa Badr and Max J. Egenhofer Hanaa Talha,: Exploring a Hybrid of Geospatial Semantic Information in Ubiquitous Computing Environments : International Journal of Computer Science Issues, Vol. 8, Issue 6, No 2 on December 2011,Page(s) 117-121
- Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. : A Practical Guide to Support Vector Classification. Page{1-16

Index Terms

Computer Science
Information Sciences

Keywords

Data Mining Information Retrieval Data Preprocessing Data Cleaning Data Warehouse