Data Quality Enhancement with String Length Distribution

Authors : Qi Xiu, Hiromu Hota, Yohsuke Ishii, Takuya Oda

Abstract : Recently, collectable manufacturing data are rapidly increasing. On the other hand, mega recall is getting serious as a social problem. Under such circumstances, there are increasing needs for preventing mega recalls by defect analysis such as root cause analysis and abnormal detection utilizing manufacturing data. However, the time to classify strings in manufacturing data by traditional method is too long to meet requirement of quick defect analysis. Therefore, we present String Length Distribution Classification method (SLDC) to correctly classify strings in a short time. This method learns character features, especially string length distribution from Product ID, Machine ID in BOM and asset list. By applying the proposal to strings in actual manufacturing data, we verified that the classification time of strings can be reduced by 80%. As a result, it can be estimated that the requirement of quick defect analysis can be fulfilled.

Keywords : string classification, data quality, feature selection, probability distribution, string length
Conference Title : ICDMTA 2017 : International Conference on Data Management Technologies and Applications
Conference Location : Madrid, Spain
Conference Dates : March 26-27, 2017

Open Science Index, Computer and Information Engineering Vol:11, No:03, 2017 publications.waset.org/abstracts/57244.pdf