Abstract

In order to deal with classification for large data, data filtering and data cleansing are used as preprocessing methods. Generally it remove noisy data, misclassified data, errors and inconsistent data and results unreliable classification. Because sometimes cleaned data can also affect the prediction accuracy or other testing. In this paper, we performed analysis of misclassified data and identify how much data has been wrong classified. For future aspect, this misclassified data is need to be rectified to get valuable information. To demonstrate this concept, we have used Air Traffic dataset from Statistical Computing Statistical Graphics (SCSG) to examine misclassified content in data set. Five supervised classifiers are used: Support vector Machine, decision procedure, k-nearest neighbor, random forest and logistic regression. The results shows that out of these classifiers, SVM classify 86% of the data correctly and only 14% of data has misclassification.

References


Index Terms

Computer Science

Distributed Systems

Keywords

Misclassification, Big Data, Classification