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.
Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV, pp. 139-168.

20. Gong, Ke, Panpan Wang, and Yi Peng. "Fault-tolerant enhanced bijective soft set with

21. O. Okun, G. Valentini, (Eds.), Supervised and Unsupervised Ensemble Methods and

22. Nodarakis, Nikolaos, Evaggelia Pitoura, Spyros Sioutas, Athanasios Tsakalidis,
Dimitrios Tsoumakos, and Giannis Tzimas. "kdann+: A rapid aknn classifier for big data."In
Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV, pp. 139-168.


24. Mood, Carina. "Logistic regression: Why we cannot do what we think we can do, and
what we can do about it." European sociological review 26, no. 1 (2010): 67-82.

Part C, Vol 1, No. 11, November Top Down Induction Of Decision Tree Classifier-A
Survey,2002


28. R. Jhawar, V. Piuri, and M. D. Santambrogio, "Fault tolerance management in IaaS
clouds." In Satellite Telecommunications (ESTEL), 2012 IEEE 1st AESS European Conference,
pp. 1-6, 2012.

29. A. Avižienis, J.C. Laprie, B. Randell, and C. Landwehr. “Basic concepts and taxonomy of
dependable and secure computing," IEEE Transactions on Dependable and Secure Computing,

group classification:Impact of predictor distribution, covariance inequality, effect size, sample


**Index Terms**

Computer Science  Distributed Systems

**Keywords**

Misclassification, Big Data, Classification