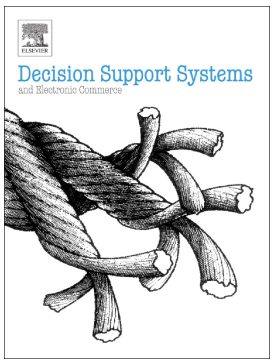

Rebuilding sample distributions for small dataset learning

Der-Chiang Li, Wu-Kuo Lin, Chien-Chih Chen, Hung-Yu Chen, Liang-Sian Lin

PII: S0167-9236(17)30201-4
DOI: doi:10.1016/j.dss.2017.10.013
Reference: DECSUP 12894

To appear in: *Decision Support Systems*

Received date: 31 March 2017
Revised date: 28 October 2017
Accepted date: 29 October 2017

Please cite this article as: Der-Chiang Li, Wu-Kuo Lin, Chien-Chih Chen, Hung-Yu Chen, Liang-Sian Lin , Rebuilding sample distributions for small dataset learning. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Decsup(2017), doi:10.1016/j.dss.2017.10.013

Rebuilding Sample Distributions for Small Dataset Learning

Der-Chiang Li*, Wu-Kuo Lin, Chien-Chih Chen, Hung-Yu Chen, Liang-Sian Lin
Department of Industrial and Information Management, National Cheng Kung University, University Road, Tainan 70101, Taiwan, R.O.C
E-mail address: *lidc@mail.ncku.edu.tw
*Corresponding author. Tel: (886) 2757575 x53134; Fax: (886) 2374252

Abstract

Over the past few decades, a few learning algorithms have been proposed to extract knowledge from data. The majority of these algorithms have been developed with the assumption that training sets can denote populations. When the training sets contain only a few properties of their populations, the algorithms may extract minimal and/or biased knowledge for decision makers. This study develops a systematic procedure based on fuzzy theories to create new training sets by rebuilding the possible sample distributions, where the procedure contains new functions that estimate domains and a sample generating method. In this study, two real cases of a leading company in the thin film transistor liquid crystal display (TFT-LCD) industry are examined. Two learning algorithms—a back-propagation neural network and support vector regression—are employed for modeling, and two sample generation approaches—bootstrap aggregating (bagging) and the synthetic minority over-sampling technique (SMOTE)—are employed to compare the accuracy of the models. The results indicate that the proposed method outperforms bagging and the SMOTE with the greatest amount of statistical support.

Keyword: Small data, virtual sample, data preprocessing