**2015 – 01**

# Influence Analysis of Binary Classification

Bo-Shiang Ke and Yuan-chin Ivan Chang

**Oct. 29, 2015**

# Influence Analysis of Binary Classification

Bo-Shiang Ke[a], Yuan-chin Ivan Chang[b,*]

[a] *Institute of Statistics, National Chiao Tung University, 1001, University Road, Hsinchu 300, Taiwan*
[b] *Institute of Statistical Science, Academia Sinica, 128, Academia Road Sec. 2, Taipei 115, Taiwan*

**Abstract**

For the sake of choosing the most appropriate classifier among a bunch of candidates, assessments of these classifiers would also play an important role in modeling; nonetheless, once the influential cases exist in some of classifiers, it is likely to make wrong decisions due to those cases. Therefore, it is necessary to identify those potential influential cases in each classifier so that the comparisons could be fair enough. In this paper, we aims to investigate those influential observations from both graphical and theoretical approaches. The graphical approach focuses on the cumulative lift chart which is frequently used in marketing and sales applications while the theoretical approaches utilize the are under ROC curve in signal detection theory in terms of influence function and local influence methods.

*Keywords:* AUC, local influence, influence function, cumulative lift chart

## 1. Introduction

Classification is one of the major topics in supervised learning, plenty of impressive classifiers (Webb and Copsey, 2011; James et al., 2013) and related performance assessments Hand (2012) have proposed from a variety of research areas and considerations. Although we have numerous metrics to quantify the performance, they may also suffer the effects caused by influential cases. It is important to identify those influential cases foe each classifier so that the comparison would be fair. Hampel (1974) proposed the influence function as a standard mathematical tool to measure the influence for cases. Cook (1986) and Wu and Luo (1993) utilized local influence in terms of slope and curvature by not only generalizing the concept of influence function but also avoiding the masking effect. Both theoretical approaches should based on an interested parameter as the objective function, this parameter is chosen by the area under ROC curve (AUC) which plays the major role in classification. Despite these

*Corresponding author
*Email address:* ycchang@stat.sinica.edu.tw (Yuan-chin Ivan Chang )