UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Stacking Ensemble Model for Liver Stiffness Classification with Imbalanced Data

by

Mingjian Wang

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Master of Biomedical Engineering by Research

Sydney, Australia

February 2021

Certificate of Authorship/Originality

I, Mingjian Wang declare that this thesis, is submitted in fulfilment of the requirements for the award of Master of Biomedical Engineering by Research, in the Faculty of Engineering and IT at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature: Signature removed prior to publication.

Date: 18/02/2021

ABSTRACT

Stacking Ensemble Model for Liver Stiffness Classification with Imbalanced Data

by

Mingjian Wang

Liver cirrhosis is a significant threat to humans; once the liver reaches the last stage of cirrhosis, there is no cure for it. Therefore, discovering cirrhosis in the early stage is one of the effective ways to decrease the mortality rate of cirrhosis. Besides early detection, increasing the correct cirrhosis diagnosis rate is another desirable method to avoid late treatment for patients. This thesis developed an automatic diagnosis approach to predict doctors' opinions for patients regarding the liver stiffness measurements from FibroScan tests.]A model using the Stacking ensemble method was presented to build a classifier for an imbalanced liver stiffness measurement data-set. The data-set was collected from 13,418 Chinese patients who had liver cirrhosis tests by FibroScan. It recorded 30 sets of features, also provided professional doctors' opinions in Chinese. To transfer the Chinese characters to digital, we applied Jieba module in Python which is a natural language processing method to create 6 labels in classification. Each label presents one doctors' opinion. Since this data-set is highly imbalanced, sampling methods such as the under-sampling method and the oversampling method are applied to solve this problem. To identify the most suitable model for the classification, we performed a study of 7 supervised learning algorithms, Logistic Regression (LR), Decision Tree (DT), Naive Bayesian (NB), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest (RF) and AdaBoost; also demonstrated the stacking models based on these supervised learning algorithms. The results demonstrated that the use of Synthetic Minority Oversampling Technique (SMOTE) oversampling technique was effective to handle the imbalanced liver data-set, and the best fitting model was constructed by using DT as meta-classifier with four base classifiers (KNN, RF, DT, SVM) in the stacking model.

Dissertation directed by Professor Steven Su

School of Biomedical Engineering

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor Steven Su for my postgraduate study. It was a great opportunity for me to work with this wonderful supervisor. I have enhanced my abilities under his helpful guidance, endless patience, consideration and kindness. Also, I would like to thank my co-supervisor Steve Ling for his helpful advice for part of my project.

Many thanks to my labmates at Faculty of Engineering and Information Technology (FEIT), for their support, giving useful recommendations and comments for my project any time I asked for help during these two years of study.

Finally, I would like to thank my parents for their encouragement. I could not achieve this study without their support.

Mingjian Wang

Sydney, Australia, 2021.

Contents

List of Figures

List of Tables