

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

© Copyright 2021 Mingjian Wang

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

Certificate	ii
Abstract	iii
Acknowledgments	iv
List of Figures	viii
List of Tables	xi
1 Introduction	1
1.1 Background	1
1.2 Problem Statement	5
1.2.1 Imbalanced Data	5
1.2.2 The Choice of Algorithms	7
1.3 Dissertation Contributions	9
1.4 Research Objectives	10
1.5 Thesis Organization	11
2 Literature Review	12
2.1 Liver Cirrhosis	12
2.2 Liver Stiffness Measurement by FibroScan	13
2.3 Imbalanced Learning Methods	17
2.3.1 Sampling Methods	19
2.3.1.1 Oversampling	19
2.3.1.1.1 Random Oversampler	19
2.3.1.1.2 Synthetic Minority Oversampling Technique (SMOTE)	20
2.3.1.2 Undersampling	21
2.3.1.2.1 EasyEnsemble	22
2.4 Natural Language Processing	23
2.4.1 The Principle of Jieba Word Segmentation	26

2.5	Classification Algorithms	27
2.5.1	Support Vector Machine	27
2.5.2	Decision Tree	33
2.5.3	K-Nearest Neighbour	35
2.5.4	Logistic Regression	38
2.5.4.1	Parameter solving	38
2.5.4.2	Regularization	39
2.5.4.3	Softmax	39
2.5.5	Naive Bayesian	41
2.6	Ensemble Methods	43
2.6.1	Bagging-Random Forest	44
2.6.2	Boosting-Adaboost	46
2.6.3	Stacked Generalisation	49
3	Methodology	52
3.1	Liver Stiffness Measurement Data-set	52
3.1.1	Data Pre-processing	52
3.2	Evaluation Criteria of Classification Performance	54
3.3	Experiment Procedure	60
4	Experiment Results and Discussion	63
4.1	Equipment	63
4.2	Experiment Results and Discussion	63
4.2.1	Simple Algorithms, Bagging and Boosting Algorithms Performances	64
4.2.1.1	KNN Algorithm	64
4.2.1.2	RF Algorithm	68
4.2.1.3	NB Algorithm	71
4.2.1.4	LR Algorithm	74
4.2.1.5	DT Algorithm	77
4.2.1.6	SVM Algorithm	80
4.2.1.7	AdaBoost Algorithm	83
4.3	Stacking Models Performances	86

4.3.1	Stacking KNN Model	86
4.3.2	Stacking RF Model	88
4.3.3	Stacking NB Model	90
4.3.4	Stacking LR Model	92
4.3.5	Stacking DT Model	94
4.3.6	Stacking SVM Model	96
4.3.7	Stacking AdaBoost Model	98
5	Conclusion and Future Work	102
5.1	Main Findings	102
5.2	Dissertation Contributions	102
5.3	Future work	104

List of Figures

2.1	Stages of liver damage (Science Source)	13
2.2	Example of imbalanced data distribution: 20 instances (16 circles belonging to Class 1 and 4 triangles belonging to Class 2) with 5:1 amount proportion	17
2.3	Decision Tree classifier decision area problem due to imbalanced data	18
2.4	Influence of noisy imbalanced data for Decision Tree classification	19
2.5	Example of new instance creation by SMOTE: the new instance $x_{i(new)}$ is located between x_i and $x_{i(n)}$	21
2.6	Original data: without any sampling process; EasyEnsemble: randomly select 50 instances from the negative class and compare with the positive class; Randomly Oversampling: repeatedly extract and generate 500 data from the positive class (which will inevitably repeat), and compare with the negative class; SMOTE: By finding the nearest neighbors of the instances in the positive class, "500 - 50 = 450 new instances in the positive class" are synthesized and merged with the original data.	22
2.7	The flow chart of Jieba principle	28
2.8	Schematic illustration of bagging, boosting and random forest (Yang et al. 2016)	44
2.9	The principle of Random Forest (Boulesteix et al. 2012)	45
2.10	Illustration of a Stacking Model with Logistic Regression and Naive Bayesian classifiers	51
3.1	Example of confusion matrix in multi-class classification tasks	57
3.2	Example of ROC curve in binary classification tasks	59
3.3	Example of a stacking model assuming RF, DT, LR, KNN and NB as base classifiers in Level-0, SVM and AdaBoost as meta-classifiers in Level-1	62

4.1	Confusion matrix and ROC curve of KNN classifier on (a) Original LSM data-set (b) EasyEnsemble undersampled LSM data-set (c) SMOTE oversampled LSM data-set	67
4.2	Confusion matrix and ROC curve of RF classifier on (a) Original LSM data-set (b) EasyEnsemble undersampled LSM data-set (c) SMOTE oversampled LSM data-set	70
4.3	Confusion matrix and ROC curve of NB classifier on (a) Original LSM data-set (b) EasyEnsemble undersampled LSM data-set (c) SMOTE oversampled LSM data-set	73
4.4	Confusion matrix and ROC curve of LR classifier on (a) Original LSM data-set (b) EasyEnsemble undersampled LSM data-set (c) SMOTE oversampled LSM data-set	76
4.5	Confusion matrix and ROC curve of DT classifier on (a) Original LSM data-set (b) EasyEnsemble undersampled LSM data-set (c) SMOTE oversampled LSM data-set	79
4.6	Confusion matrix and ROC curve of SVM classifier on (a) Original LSM data-set (b) EasyEnsemble undersampled LSM data-set (c) SMOTE oversampled LSM data-set	82
4.7	Confusion matrix and ROC curve of AdaBoost classifier on (a) Original LSM data-set (b) EasyEnsemble undersampled LSM data-set (c) SMOTE oversampled LSM data-set	85
4.8	Confusion matrix and ROC curve of the stacking KNN model on (a) Original data-set (b) SMOTE oversampled data-set	88
4.9	Confusion matrix and ROC curve of the stacking RF model on (a) Original data-set (b) SMOTE oversampled data-set	90
4.10	Confusion matrix and ROC curve of the stacking NB model on (a) Original data-set (b) SMOTE oversampled data-set	92
4.11	Confusion matrix and ROC curve of the stacking LR model on (a) Original data-set (b) SMOTE oversampled data-set	94
4.12	Confusion matrix and ROC curve of the stacking DT model on (a) Original data-set (b) SMOTE oversampled data-set	96
4.13	Confusion matrix and ROC curve of the stacking SVM model on (a) Original data-set (b) SMOTE oversampled data-set	98
4.14	Confusion matrix and ROC curve of the stacking AdaBoost model on (a) Original data-set (b) SMOTE oversampled data-set	100

5.1 Proposed stacking model (KNN, RF, DT, SVM as base classifiers and RF as meta-classifier)	104
--	-----

List of Tables

2.1	Possible factors effecting the accuracy of FibroScan results	15
2.2	Various conditions of liver disease with different cutoff values: F_0 to F_1 is the first stage of liver fibrosis; F_2 is the second stage of liver fibrosis; F_3 to F_4 is the stage of liver cirrhosis (University of Health Network 2018)	16
2.3	Advantages and disadvantages of Support Vector Machine	32
2.4	Advantages and disadvantages of Decision Tree	35
2.5	Advantages and disadvantages of K-Nearest Neighbour	37
2.6	Advantages and disadvantages of Logistic Regression	40
2.7	Advantages and disadvantages of Naive Bayesian	43
2.8	Advantages and disadvantages of Random Forest	46
2.9	Advantages and disadvantages of AdaBoost	49
3.1	Data-set variables description	53
3.2	Second opinion in Chinese character and the output of Jieba module in Python	53
3.3	The mean and variance values of part of features of liver stiffness measurement data-set	54
3.4	Confusion matrix of binary classification tasks	55
4.1	Performance obtained by KNN classifier on original LSM data-set	65
4.2	Performance obtained by KNN classifier on undersampled LSM data-set	66
4.3	Performance obtained by KNN classifier on oversampled LSM data-set	66
4.4	Performance obtained by RF classifier on original LSM data-set	68
4.5	Performance obtained by RF classifier on undersampled LSM data-set	69
4.6	Performance obtained by RF classifier on oversampled LSM data-set	69
4.7	Performance obtained by NB classifier on original LSM data-set	71
4.8	Performance obtained by NB classifier on undersampled LSM data-set	72

4.9	Performance obtained by NB classifier on oversampled LSM data-set	72
4.10	Performance obtained by LR classifier on original LSM data-set	74
4.11	Performance obtained by LR classifier on undersampled LSM data-set . . .	75
4.12	Performance obtained by LR classifier on oversampled LSM data-set	75
4.13	Performance obtained by DT classifier on original LSM data-set	77
4.14	Performance obtained by DT classifier on undersampled LSM data-set . . .	78
4.15	Performance obtained by DT classifier on oversampled LSM data-set	78
4.16	Performance obtained by SVM classifier on original LSM data-set	80
4.17	Performance obtained by SVM classifier on undersampled LSM data-set . .	81
4.18	Performance obtained by SVM classifier on oversampled LSM data-set . . .	81
4.19	Performance obtained by AdaBoost classifier on original LSM data-set . . .	83
4.20	Performance obtained by AdaBoost classifier on undersampled LSM data-set	84
4.21	Performance obtained by AdaBoost classifier on oversampled LSM data-set	84
4.22	Performance obtained by the stacking KNN model on original data-set . . .	87
4.23	Performance obtained by the stacking KNN model on SMOTE oversampled data-set	87
4.24	Performance obtained by the stacking RF model on original data-set	89
4.25	Performance obtained by the stacking RF model on SMOTE oversampled data-set	89
4.26	Performance obtained by the stacking NB model on original data-set	91
4.27	Performance obtained by the stacking NB model on SMOTE oversampled data-set	91
4.28	Performance obtained by the stacking LR model on original data-set	93
4.29	Performance obtained by the stacking LR classifier on SMOTE oversampled data-set	93
4.30	Performance obtained by the stacking DT model on original data-set	95
4.31	Performance obtained by the stacking DT model on SMOTE oversampled data-set	95
4.32	Performance obtained by the stacking SVM model on original data-set . . .	97
4.33	Performance obtained by the stacking SVM model on oversampled data-set	97
4.34	Performance obtained by the stacking AdaBoost model on original data-set	99

4.35 Performance obtained by the stacking AdaBoost model on SMOTE oversampled data-set	99
4.36 Performance obtained by selected three stacking models on SMOTE oversampled data-set	101