Structure Type Classification Based on Tax Assessor Data

Yue (Major) Zeng

majorzeng@Stanford.edu

Stanford
CS229 Project

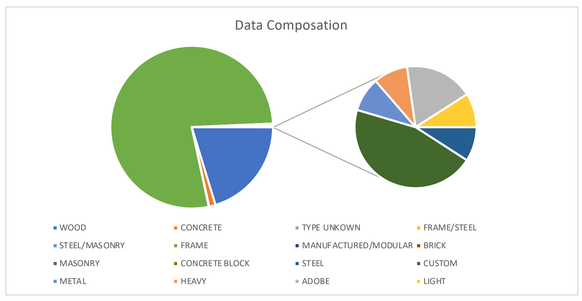
Motivation

In order to predict regional damage of earthquakes, an important input information is the structural type of each building in the local building stock, which is a strong indicator of the strength of a building. In situ assessment of a professional engineer provides the most accurate information, however, it is often infeasible due to the shear quantity of buildings existing in urban areas. This project intends to use machine learning tools to explore a set of tax assessor data and make predictions on structural types.

Data

The data used in this study is tax assessor file of San Mateo county for the year of 2016. It includes 128137 examples and 149 features. The dataset is sparse. Feature ”construction type” the labels to predict and its only 18% filled. All categorical data are transferred into numerical form by one hot encoding, including the label”. The result data frame contains 23143 labeled examples with 531 features each and 16 mutually exclusive binary labels.

Imbalanced Data Handling



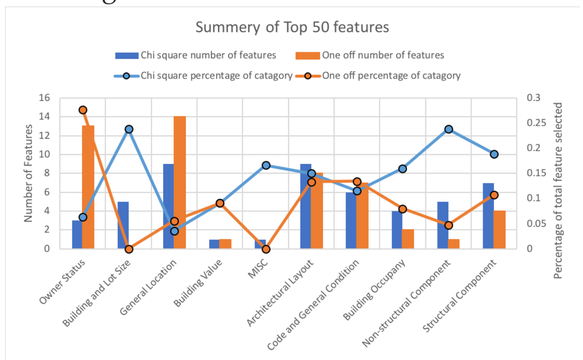
Three methods are considered here to balance the data: **1) random over sampling, 2) random under sampling and 3) adding class weights**. A simple binary relevance classifier with logistic regression is used to compare there performance. The evaluating criteria are general accuracy and number of examples misclassified in rare classes including both false positive and false negative.

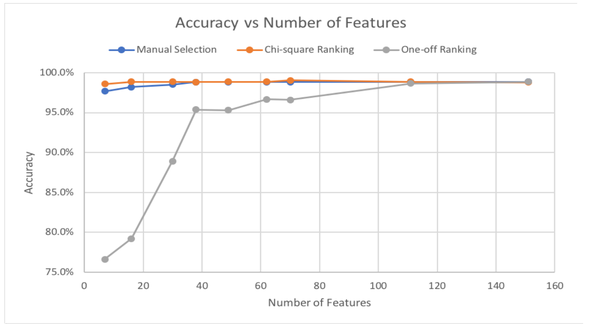
		Overall Accuracy	Number of Misclassified in Rare Classes
Baseline	No Manipulation	98.6%	49
Method (1)	Random Over Exampling	96.3%	104
Method (2)	Random Under Exampling	94.1%	140
Method (3)	Train with Weighted Examples	96.3%	106

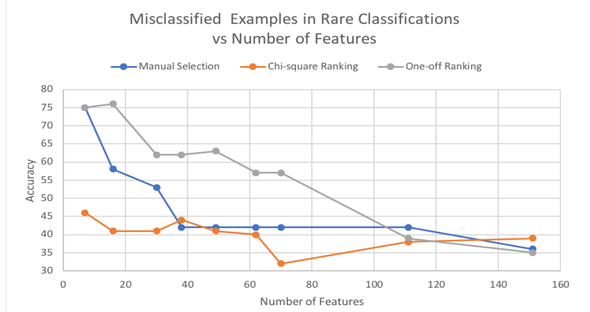
The result shows that all three methods are not effective, performing worse than baseline in both criteria.

Feature Selection Filters

The many features exists in the data set and expensive to collect. Three methods are considered here to rank the features: **1) chi-square score, 2) one off testing** and 3) manual selection. The result shows that Chi-square is the best ranking mechanism. Ranking shows that location and some physical characters of the building are informative. The value of buildings or owner status are less relevant.

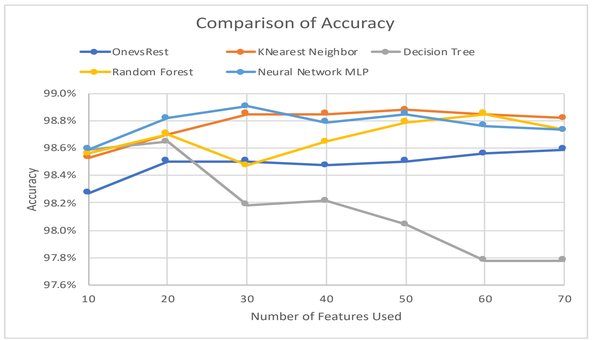


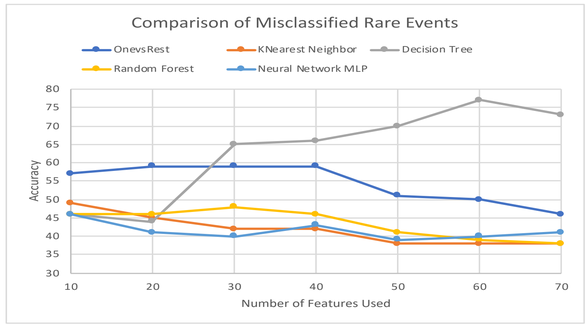




Classification Models and Number of Features to use

Five multilabel classification methods are tested here with various number of features: **1) binary relevance classifier with logistic regression, 2) k-nearest neighbor classifier, 3) decision tree classifier, 4) random forest classifier and 5) 2 layer neural network with logistic activation.**





Result shows that k-nearest neighbor and neural network are the better classification models, with more steady performance as the number of feature decrease.

Prediction and Discussion

Prediction on the unlabeled data shows majority (96%~98%) of examples classified in the common classes. Due to the sparseness of the data and potential inaccuracy in collection process, it’s not possible to evaluate the accuracy such prediction. The limited results in this study shows that more work is needed in order to predict structural type with reasonable confident.

Future Work

In order to better classify rare classes, one could collect more complete data sets, especially more examples of the rare classes or includes more advanced forms of penalty function in optimization to emphasis the rare classes. Also searching for cases where regional statistics are available for percentage of building types could validate the prediction results. Finally, more case studies in feature selection could yield useful guidelines for targeted feature collection.