Analysis of Data Sampling Methods in In-flight Load Monitoring Model

Yeon-woo Shin

June 2022

1 Abstract

As autonomous flight system has been rising research topic, machine learning approach on such topic has gained great attention. There are wide range of applications of machine learning to improve the flight. For instance, the mechanical damages inside of flight engine can be detected by novel anomaly detection algorithm such as Support Vector Machine. The report introduces a flight maneuver classification model and data sampling methods that were used to address class imbalance issue. Also, the problems in the applied data sampling methods are explicitly described, and suggested solutions were provided for future improvement.

2 Summary

A flight maneuver classification model is to decide which state the aircraft is in out of following: Cruising, Ascending, Descending, Left turn, Right turn based on data from FBG Strain sensor.

2.1 Form Dataset

There were 12 sensors in total, and the graph of measured strain varying over time is drawn for each sensor. For each graph, the slicing window is applied to record the strain over 1 second. The sliced data for all 12 sensors over equivalent time frame then get horizontally connected in the dataset.

2.2 Pre-process Dataset

To catch up the speed at action state of aircraft alters, predicting the action of the aircraft based on strain dataset has to be fast. To reduce the computational time, PCA (Principal Component Analysis) was used. PCA is unsupervised machine learning technique to reduce the dimensions of dataset by finding the optimal principal axis that can re-map the data in a way that the maximum variance between features are preserved. Even though Explained Variance Ratio(EVR) for PCA was over 99%, the valid reason for chosen number of principal components wasn't specifically given. The report should have provided the graph such shown in Figure [2] to

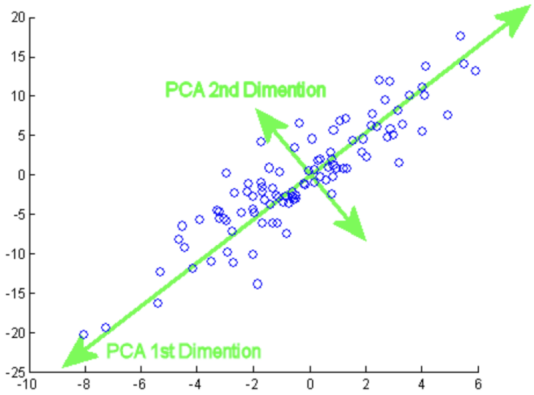

Figure 1: Principal Axis for Sample Dataset

justify their selection in the optimal number of principal components. Figure [2] implies that increasing the number of components will not contribute to maximizing explained variance ratio anymore, so there might be useless surplus that will simply increase computational complexity if higher number of components are used. To avoid having surplus components, it's important to carefully observe at the growth of explained variance ratio over number of components and find the optimal number of principal components which guarantees maximized variance and are the most computationally efficient.

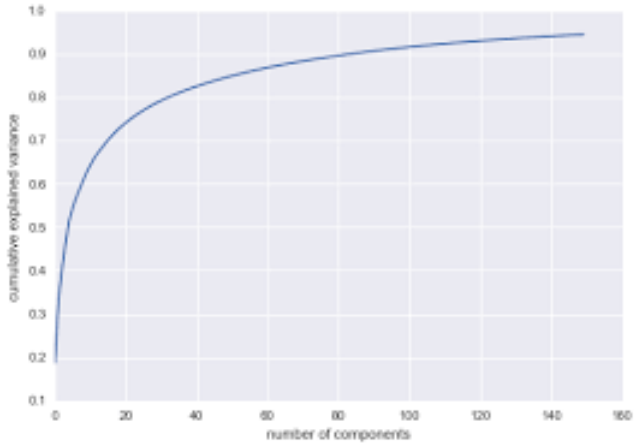

Figure 2: Graph of explained variance ratio over number of principal components

Also, to address class imbalance issue, SMOTE and cost-sensitive method were used. SMOTE is over-sampling method that generates new synthetic samples using the original data. Cost-sensitive method is to increase the penalty for misclassifying the minority samples to prevent all sample inputs being classified into majority class.

2.3 Build a model

The dataset balanced by SMOTE was then applied to linear classification models as following: Logistic Regression, Suuport Vector Machine, etc.

3 Problem

Due to nonequivalent probability of occurrence of classes, imbalanced dataset becomes one of the main problems of real-world classification task. The report proposes two following methods to solve an issue caused by imbalanced dataset: SMOTE and cost-sensitive method. However, there are significant drawbacks of those methods.

SMOTE

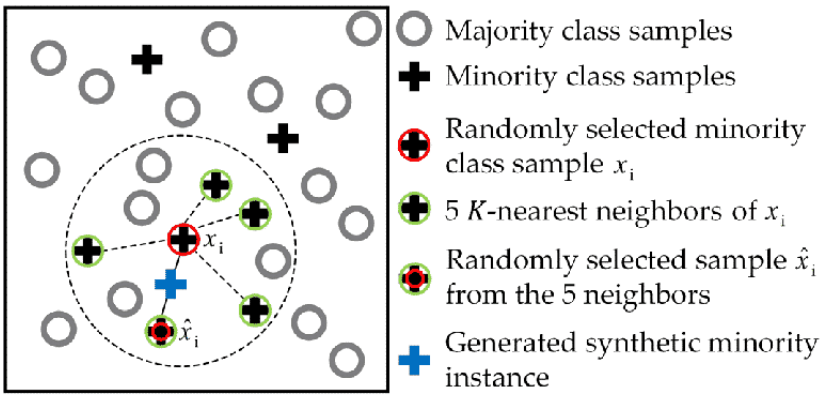

Figure 3: Example of how SMOTE generate new sample

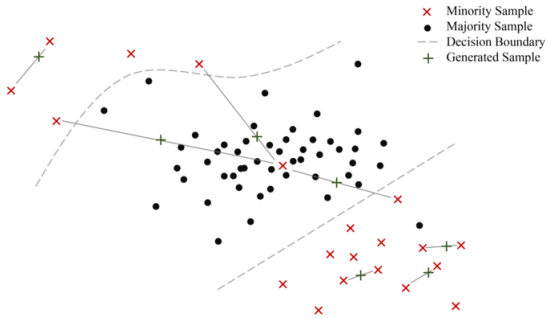

Figure 4: Example of failure of SMOTE to generate noise-resistant samples

Abbreviated from Synthetic Minority Oversampling Technique, SMOTE is widely used since it generates new samples by average the randomly selected minority class sample x and one of nearest neighbors of x instead of simply duplicating the existing sample in the minority class, thereby addressing the overfitting problem. As shown in Figure 3, SMOTE finds the midpoint between minority sample and one of its nearest neighbors and set it as the new sample. However, the problem of SMOTE is its algorithmic weakness against noise or skew. SMOTE could use the outliers in the dataset to generate new samples as shown in Figure 4; SMOTE is very insensitive to the existing noise and could therefore amplify it by generating new samples, rather hindering the classification capacity.

Cost-sensitive Learning

While SMOTE was data-level solution for imbalanced dataset since it oversamples the minority data, Cost-sensitive learning is algorithm-level solution because it's applied to the classifier to minimize the effect of bias toward the majority dataset. The problem with cost-sensitive method is that there is no mathematically strict approach to find the value of extra penalty given to the model when it doesn't classify minority class sample correct.

4 Solutions

4.1 BMPM (Biased Minimax Probability Machine)

Instead of increasing the penalty for misclassifying the data in minority class, as suggested in the paper, BMPM increases the accuracy of classifying the data in minority class while setting the accuracy of classifying the data in majority class in acceptable level.

$$\max_{\alpha,\beta,b,a\neq 0} = \alpha \tag{1}$$

$$\inf_{x\sim(x,\Sigma \mathrm{x})} \Pr(a^T x \geq b) \geq \alpha \tag{2}$$

$$\inf_{y\sim(y,\Sigma \mathrm{y})} \Pr(a^T y \leq b) \geq \beta \tag{3}$$

$$\beta \geq \beta_0, \tag{4}$$

β_0 is minimal acceptable value of accuracy for classification of majority class. The equations above basically are to maximize α, which stands for accuracy for classification of minority class. BMPM provides systematic approach to control the accuracy of classification of majority and minority class. BMPM gives an classifying bias to the minority class while keeping the bias in appropriate range so that the classification accuracy of data in majority class is guaranteed to be in acceptable range.

4.2 Balanced Bagging Classifier

Another algorithm-level solution to address the problem arising from an imbalanced dataset could be balanced bagging classifier. Bagging classifier is the model that ensembles sub-models, each of them

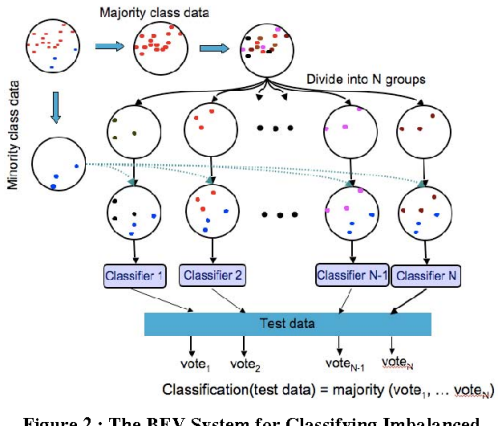

Figure 5: Mechanism of Balanced Bagging Classifier

trained by subset of training dataset. To address the problem of oversampling solution, undersampling could be used as an alternative. However, undersampling might not be efficient in terms of training the model because as literally meant in its name, the training samples are reduced. To effectively use the training samples while executing the undersampling, balanced bagging classifier should be applied. Instead of training a model with a single set of minority-class dataset and reduced majority-class dataset, the several balanced sets of divided majority-dataset and the minority-class dataset could be used to train different models. To formulate the final prediction using the predictions made by several models, majority voting ensemble method is used.

5 Reference

This report was written under internship work in analysis of AI algorithm for DIGITLOG Inc.