Analysis of Density between Classes for Mitigating Class Imbalance Problem for Activity Recognition

Sayeda Shamma Alia† Sozo Inoue‡

Kyushu Institute of Technology† Kyushu Institute of Technology‡

Abstract

Class Imbalance problem is unavoidable problem for activity recognition using mobile sensors in real life scenario when we have very less amount of data for some activity classes. It can affect the accuracy of the algorithms for classification. In this paper we measure the impact of class density in the accuracy of classification in imbalance cases. It is important for understanding the problem better that can help in finding better solution for classification in this scenario. Our initial experiment shows that- class imbalance affects the performance of classifier negatively and the higher the value of density (lower deviation), the better the performance of the classifier becomes.

Introduction

Automatically recognizing activities performed by human is gaining more attention day by day [1]. A data set is "imbalanced" if its dependent variable is categorical and the number of instances in one class is different from those in the other class. All the activity classes do not have same number of instances because it is practically impossible for a human to spend same amount of time for every activity he performs. If we use sliding windows for segmentation, activities of longer duration will have more samples (e.g.: sleeping will have more windows than eating). For these reasons naturally imbalance occurs in activity recognition data. Sample size for each class plays a vital role in classification [2]. Also, sometimes there are some rare or unusual activities that can happen very occasionally (like: Heart attack, Slipping, Robbery etc.). Thus, there will be very less amount of data of these activities. With few data traditional classifier fails to recognize these activities properly. To predict an activity class accurately it is important to learn perfectly form these imbalanced classes. As class imbalance is bound to occur in activity recognition, it is better to handle this problem for a better prediction of the activity class. Misclassifying imbalanced class can result in heavy cost or harmful result. If we can properly detect the imbalanced class, it can lessen the level of harm (by detecting heart attack it can prevent some worst case scenarios). In some cases it is very important to detect the imbalanced classes accurately while avoiding false positives, for example: in the heart attack scenario, the intention is to detect every heart attack without creating many false alarms. Hence, imbalanced class should be classified correctly. If density of the imbalanced class is higher which means lower deviation, the prediction rate of the imbalanced class can improve. Here, by "density" we implied diversification of deviation. If the deviation is high, all the instances will be scattered, which indicates lower density. And, if the deviation is low, all the samples will be near to each other, which implies higher density.

Based on the aforementioned discussion, we tried to find out the impact of density on classification and answer the following research questions:

- If there is any impact of deviation on classification of imbalanced data
    - For If the impact is positive or negative on the performance

There are existing works [3, 4] that tried to diminish the class imbalance problem in activity recognition. [3] used ensemble classifiers (like- Bagging, Boosting) for classification. Besides, resampling methods are used in the existing literature [4]. The main goal of this work is to find out the effect of density (deviation) on classification for imbalanced class. The contributions of this work is summarized as follows

- We established a relationship between deviation of a class and performance of classifier.
- For balanced data and same deviation the accuracy was 93% and for lower deviation it was 98%. For imbalanced data the accuracy was 96% and 97% respectively. Therefore, lower deviation can contribute positively in prediction of the class.

Experimental Settings

The hypothesis of this work is that- "*if the density of the imbalanced class is higher, it can perform better. More specifically if the deviation of the class is reduced, consequently the performance will improve*".

To evaluate the hypothesis, we applied the hypothesis on two datasets described below.

- Synthetic dataset that follows Gaussian distribution. This dataset has no noise while the HAR data set has noise. There are two classes (class A and class B) and one feature. There were 900 instances for each class for balanced data and 100 and 900 instances respectively for Class A and B for imbalanced data.

† 「Sayeda Shamma Alia・ Kyushu Institute of Technology」

‡ 「Sozo Inoue・ Kyushu Institute of Technology」

- For Human Activity Recognition Using Smartphones Data Set[1] (HAR). There are six classes (Walking, walking upstairs, walking downstairs, sitting, standing and laying) and 561 features, but for this experiment we used Principal component analysis for feature selection and used 161 features. The original HAR dataset is fairly balanced. The number of instances in each activity class of the HAR dataset is showed in Table 1. On that account, sample size of "walking" class was reduced to make the dataset imbalanced.

Table 1: Sample size in HAR data

Activities	Number of Instances
Walking	1226
Walking Upstairs	1073
Waking Downstairs	986
Standing	1367
Sitting	1286
Laying	1407

Result & Discussion

For classifying the data we used kNN (k-nearest neighbors) algorithm where k=10. Figure 1 shows the accuracies of this experiment on synthetic data. For balanced data, we can clearly see the improvement in performance when the deviation is reduced, whereas for imbalanced data, there is a little improvement in accuracy due to the change in deviation. Note that the improvement in performance can be understandable if we see the precision and recall value.

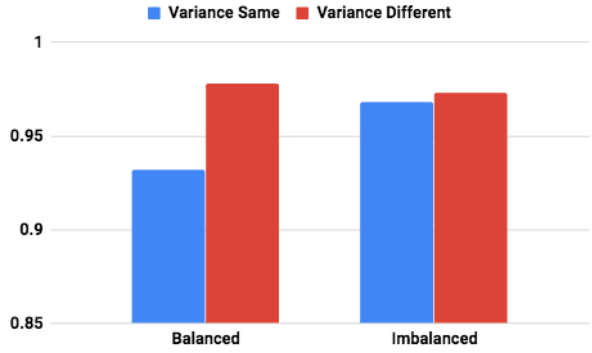

Figure 1: Accuracy comparison of Synthetic data

We saw that when the deviation is same, the precision and recall for class A is 93% for balanced data, but when the deviation is reduced for class A, the precision is 95% and recall is 100%. For imbalanced data, and same deviation precision and recall was respectively 87% and 80% and for reduced deviation it was 81% and 96% respectively. This complies with our hypothesis. For further validation the hypothesis, we conducted the same experiment on HAR dataset as well.

For this experiment we selected 1, 3, 5, 10, 15, 20 and 100% data of the "walking" class to make it imbalanced which contains (13, 37, 62, 123, 184, 246, and 1226) instances respectively and considered all the data for other classes. Figure 2 shows that the accuracy of the walking class is very low when the dataset is imbalanced and gradually improves when we incorporate more data to make it balanced. Note that, we cannot observe the effect of deviation here because deviation remains almost similar for different imbalanced data size of "walking" class used in this experiment.

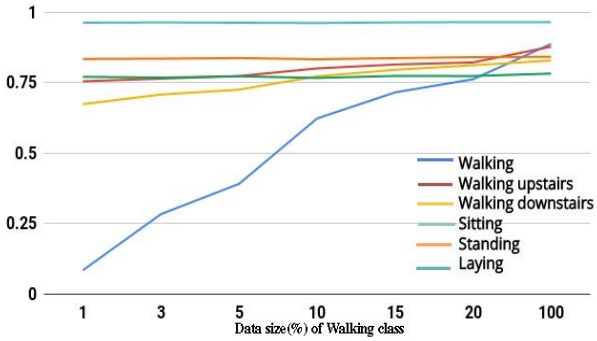

Figure 2: Accuracy of HAR data

Conclusion

In this paper we have demonstrated that the imbalanced class has impact on the classification accuracy. The strategy to fix the imbalances within class instances made an improvement which is demonstrated here empirically using synthetic data and thus proves our hypothesis. However, as there is no deviation change both in balanced and imbalanced in HAR dataset we only show the effect of imbalanced data (without the effect of deviation) here. In future we will address this using realistic data.

References

[1] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics*, 2012.

[2] T. Plotz, Yu Guan, "Ensembles of Deep LSTM Learners for Activity Recognition using Wearables," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2017.

[3] N. Dandekar, P. Mysore and M. L. Littman Nishkam Ravi, "Activity Recognition from Accelerometer Data," in *American Association for Artificial Intelligence* , 2005.

[4] M. Zeng, P. Tague, J. Zhang Le T. Nguyen, "Recognizing New Activities with Limited Training Data," in *ACM International Symposium on Wearable Computers*, 2015.

[1] https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones