CHAPTER 5

DISCUSSION AND CONCLUSION

To deal with a dataset with the imbalanced between-class distribution, performing the data pre-processing techniques to change the class distribution to become balanced is one of effective approaches chosen by researchers. Balanced datasets from these techniques can practically improve the accuracy performance of class imbalance problem. In this dissertation, the existing oversampling techniques, SMOTE and safe-level SMOTE are investigated and the framework which could enhance SMOTE in various aspects are introduced.

Three issues of SMOTE are

1. Dealing with minority outcast instances,
2. Relocating the position of conflicted synthetic instances
3. Avoiding one parameter k_n for all instances.

In order to overcome these issues, one additional process and two new oversampling techniques are introduced to be used on an imbalanced dataset. To deal with minority outcast instances, **minority outcast handling process** has been introduced. This process detects minority outcast instances based on the defined criterion. For this work, a positive instance which contains all of their c-nearest neighbors as negative is defined as a minority outcast and it is removed from the training set which is used to create synthetic instances. After the classification phase is finished with the trained model, these minority outcasts are added to build 1-nearest neighbor model with existing negative instances. The resulting model is applied with the test set resulting in increasing the chance of predicting positive. The result of this process increases the classification performance on oversampling techniques which are attached with this minority outcast handling process. As confirmed by Wilcoxon signed-rank test, the difference between an oversampling technique with the minority outcast handling process and the one without it is significantly improved.

For relocating the position of conflicted synthetic instances, a new oversampling technique called **relocating safe-level SMOTE (RSLS)** is introduced. It is a combination of the relocating process, safe-level SMOTE and the minority outcast handling. This technique is an improved technique of safe-level SMOTE. RSLS adds an additional process to relocate a synthetic instance once it is found generating near a negative instance. This technique is equipped with the minority

outcast handling which helps utilizing positive instances with the zero safe-level value. The performance of RSLS is shown through experiments and compared with other oversampling techniques. The experimental results show that RSLS has the highest number of cases which it achieves the best average F-measure. The number of datasets which RSLS has the best F-measure is consistent through classifiers. This implies that the choice of classifier does not affect the performance. Since RSLS concerns about negative instances, it still works well with the dataset which contains a few negative instances inside a group of positive instances.

The last approach is determining the parameter k. **Adaptive neighbors SMOTE (ANS)** provides the concept of assigning the parameter k_i automatically to each positive instance p_i based on its density of surrounding positive instances. With the varying values of k, synthetic instances generated from ANS scatter around the region of positive instances and represent the actual shape of overall positive instances. Most comparisons of ANS against other oversampling techniques shows that this idea has a better or equal classification performance with SMOTE (with a default fixed k = 5) and other oversampling techniques. To enhance the performance, the minority outcast handling process with 1-NN is also added. The result shows that ANS with minority outcast handling can outperform other oversampling techniques in most cases and the difference made by ANS is significantly positive after testing with the Wilcoxon sign ranked test. Since ANS does not concern about negative instances and uses the density of positive instances on assigning the parameter k, ANS might not be effective with the dataset which is not well-separated between positive and negative instances.

Future works

From the introduction of two different new oversampling techniques, there are several aspects which can be extended such as replacing c-nearest neighbor using for detecting minority outcasts with some parameter-free outlier detection ideas, or combining these two oversampling techniques into a new oversampling technique as a modified version of safe-level SMOTE with one less parameter since the number of k is already determined with the adaptive neighbor concept. However, the problem about finding the appropriate value of c still opens in this combined technique. Even changing the criterion to select minority outcast into some outlier detection ideas cannot avoid this issue since c-nearest neighbor is the important part in safe-level SMOTE as the part to compute the safe-level value.

There should be some criteria to decide the value of c in order to make this oversampling technique becomes parameter-free.