CHAPTER 5

Conclusion

From the concept mentioned in chapter 1 where we want to train supervised learning systems with large training data, this thesis is able to realize the objective by introducing a preprocessing technique to eliminate redundant and unnecessary samples from the training data.

We considered the type of data reduction for supervised learning in two types which are instance reduction for regression and for classification. A general property of regression data is that their output values are continuous. Examples of such data are temperature data, water level data, electrical power usage data, or even electrocardiograph signals as mentioned in chapter 4.

For regression, the proposed instance reduction starts by splitting a data set into smaller parts to help reduce computation cost. That is the processing time becomes faster with less memory consumption. After that, each smaller part is quantized its output data into groups. Each group is then clustered according to the input instances. The sample mean of each group is computed so as to be used as the group representative. We control errors and the number of training data by the parameter q and m as in experiment 4.1. In section 4.1.5, the ECG reconstruction plot and the RMSE values were provided and they were so close to the original ones without data reduction. Therefore, it supports that the proposed method has a property to retain the significant data.

For classification, the outputs of training data are generally discrete. They are the labels of the samples such as in experiment 4.2.3 to 4.2.5. The concept of the proposed instance reduction for classification is that the significant data are near the decision boundary and must be retained. Other data far from the decision boundary are removed as it is considered unimportant or redundant.

The reduction process for classification starts with reducing the overlap by applying the 1NN for every data class. After that, we approximate the boundary of each class by the minimum distance from the enemy classes to reduce the number of instances. This, in turn, helps reduce computing time and memory. Next, we apply some mixed search strategies, including the incremental and decremental search, to remove unimportant data and later prune the reduced data as the last step.

The proposed data reduction method is considered a hybrid type which combines the condensation and edition selection. Following the property of condensation selection, this method can reduce the training data but cannot improve the accuracy beyond the value provided by the original data. This is seen from the experimental result in section 4.2. Although the accuracy of the proposed method reported in section 4.2.3 to 4.2.5 were less than of some hybrid-wrapper based methods, the proposed method is not overly complex. In fact, it even consumed less memory and computing time than other methods.

We compared the proposed method with other methods that are based on the same type of selection (condensation and edition) as in section 4.2.3 to 4.2.5. The overall results were better than the other methods except for the RNN method whose both accuracy and reduction ratio were better. However, the computational cost for RNN method was so expensive that it could not be used to process large training data. Therefore, we omitted this method in some experiments. From the experimental results in section 4.2.6, the proposed method could reduce the training data for classification tasks. However, the accuracy obtained from the supervised learning classifier was lower than from the nearest neighbor classifier.

For further works, one possible improvement to the proposed method in the data reduction for regression problem is to develop an automatic selection algorithm to determine the parameters *m* and *q* which give the best performance.