Abstract of Main Thesis

Title of Thesis

Learning Algorithms for Classification in Imbalanced Data Environments

Phonetically in Japanese Hiragana　ぐえん　まいん　ひえん
Name:　NGUYEN MANH HIEN

Abstract on the Content of the Applicant's Thesis

Conventional learning algorithms often cannot provide satisfactory performance when classifying imbalanced data, because they assume a balanced class distribution underlying the data. This dissertation deals with the class imbalance problem in both static and dynamic learning environments. In static environments, the whole training set is known in advance, and the construction of a classification model is performed only once. To deal with class imbalance, this dissertation proposes a new over-sampling method that can help to better refine the decision boundary of support vector machines (SVM). The underlying idea is a novel strategy of over-sampling based on both interpolation and extrapolation, each of which may be more appropriate for specific locations of the borderline minority instances. The experimental results on some benchmark UCI data sets indicate that the method achieves better performance than standard SVM as well as other over-sampling methods.

In dynamic environments, however, the training instances continuously arrive in the form of data streams, and therefore imposing challenges that do not exist in static environments, such as the need to incrementally update a model with new training data, and to deal with concept drift (i.e., the data distribution is changing over time). This dissertation proposes several new methods, including (1) two online sampling methods that can incrementally learn from imbalanced data streams, and (2) two methods that can effectively handle data streams with both concept drift and class imbalance, based on reusing some of the past minority instances. The experimental results on simulated as well as real-world data streams confirm the improvement of our methods over previous methods. In addition, this dissertation conducts a comparative study of sampling methods on imbalanced data streams, from which a number of new and useful insights are drawn, such as the influence of training set size on the relative performance of sampling methods.