Abstract

In supervised machine learning, a training set of examples which are assigned to the correct target labels is a necessary prerequisite. However, in many applications, the task of assigning target labels cannot be conducted in an automatic manner, but involves human decisions and is therefore time-consuming and expensive.

In this final task, active learning is implemented in a support vector machine and examined what factors affect the amount of labeled training data and the accuracy of the system, and how they affect. It also compared the selection method of initial data and next data, the random method and the dissimilarity method. Data used in this final task is the Wisconsin Breast Cancer Diagnosis and Hill-Valley from the UCI Repository. The main goal of active learning is to select the data that is important or have influence in the system, so that it can reduce the amount of data that need to be labeled.

The results showed that active learning can reduce the amount of data need tobe labeled up to 82.5% without any significant decrease in the system accuracy.

Keywords: *Active Learning, Support Vector Machine, classification, data label, reduction.*