**使用樣本選擇降低**

**噪音資料對文章分類的影響**

**Using Instance Selection to Reduce**

**the Impact of Noise Data on Text Classification**

**陳耀輝, 孫立洋**

**Yaw-Huei Chen, Li-Yang Sun**

國立嘉義大學資訊工程學系

Department of Computer Science and Information Engineering National Chiayi University

{ychen, s1060414}@mail.ncyu.edu.tw

**摘要**

訓練一個深度學習的模型需要大量的訓練資料，但實際收集到的大量資料中不免有噪音資料，這些訓練資料中的噪音會嚴重影響模型分類的正確性。我們提出了使用多階段的樣本選擇方法來增強模型抵抗噪音訓練資料的能力，這個方法在每個階段都要挑選上階段分類正確與錯誤的兩種資料當成下階段的訓練資料。挑選分類正確的資料可以留下大部分的乾淨資料，而挑選少數上階段分類錯誤的資料可以增加訓練資料的數量與多樣性。實驗結果顯示我們的方法在含有不同比例噪音的IMDb資料集中都可以增加模型的強健性，進而達到降低噪音訓練資料對模型影響的目的。

關鍵字: 文章分類; 噪音資料; 強健訓練; 樣本選擇

**Abstract**

Training a deep learning model requires a large amount of training data, but the actual collected data may contain noisy data, and the noise in these training data can seriously affect the correctness of model classification. We propose a multi-stage instance selection method to enhance the model's ability to resist noise training data, which selects the two types of data classified correctly and incorrectly at each stage as the training data for the next stage. Picking the correctly classified data leaves most of the clean data, while selecting a small number of misclassified data from the last stage can increase the number and diversity of training data. The experimental results show that our method can increase the robustness of the model in the IMDb data set containing different proportions of noise, so as to reduce the influence of noisy training data on the model.

Keywords: text classification; noise data; robust training; instance selection