Log Data Analysis & Software Diagnosis
Centre for Mechatronics and Hybrid Technology
McMaster University
Elliott (Yixin) Huangfu

BACKGROUND

Logs are machine data generated constantly by operating system, recording software running events and status. They provide rich information for developers to track issues and reproduce errors.

The goal of this project is to detect software defects automatically by analyzing system log files.

ANALYSIS of NOMINAL DATA

Unlike numerical data from most engineering applications, log data are mostly nominal and symbolic. This means mathematical manipulation cannot be applied.

Still, the order information of log messages can be utilized as it contains valuable clues explaining software behaviours.

PATTERN DISCOVERY with DATA MINING

Logs lines are tokenized and converted into sequences of nominal values. The intuition is to extract unique sequences that appear in erroneous log sequences, while not present in regular ones.

Such approach is called contrast data mining. A prefix tree based search algorithm is implemented, and outputs a set of sequence patterns for every defect.

NAÏVE BAYES CLASSIFIER

Naïve Bayes classifier is a highly practical method in machine learning to predict target values. To apply this approach, every pattern ($p_i$) is treated as a feature, and a series of hypothesis $h_1, h_2, ... h_n$ represents certain defects happening. The conditional probability $P(p_i|h_j)$ is given in a matrix form.

During testing, a new log instance is examined to determine the occurrence of each feature pattern. Then, either $P(p_i|h_j)$ or $P(\neg p_i|h_j)$ will be selected for the calculation. Naïve Bayes Classifier outputs the hypothesis with largest probability product.

RESULT & NEXT-UPS

The method has achieved overall 72% accuracy, and shows good rejection against noise data, which means a low possibility of giving false positives.

However, 37% accuracy in the valid group shows potential for improvement. Future works will focus on enhancing the quality of feature extraction with recurrent neural networks.