Kernel Methods Data Challenge Report

Mohamed Amine Rguig
Université Paris-Dauphine

Introduction

Graphs are commonly used to model complex relationships in scientific and engineering fields, and graph classification is important for predicting unknown graphs or understanding structures among different categories. The kernel method is a powerful solution for graph classification, but it ignores feature selection and sub-structures with low discriminative power can decrease accuracy. To solve this, We followed [1] an efficient graph classification algorithm based on graph set reconstruction and graph kernel feature reduction .

1 Graph Set Reconstruction

We employed the gSPAN algorithm to mine frequent subgraphs, using the code provided by LASSEREGIN in [2]. After successfully mining the frequent subgraphs, we proceeded to compute their discriminative score, experimenting with different threshold values s. Unfortunately, we were unable to identify any discriminant subgraphs with a score higher than 0.5. As a result, we selected the least discriminant subgraphs with a score lower than 0.3 and the most discriminant subgraphs with a score higher than 0.3 for further analysis. Additionally, we removed nodes from LDFS if they did not overlap with MDFS in each graph. While the paper suggested including infrequent subgraphs as edges after removing the frequent ones, we ultimately decided against doing so. This was due to the fact that most subgraphs had a discriminative score that did not surpass 0.15, which led us to fear the possibility of information loss.

2 Graph kernels

To compute kernels, we utilized the GRAKEL library and specifically opted to use the Weisfeiler-Lehman Subtree kernel and Weisfeiler-Lehman Optimal Assignment kernel. We selected these kernels due to their efficient computation time and their ability to express information effectively across many well-known datasets. After experimenting with both kernels, we determined that the Weisfeiler-Lehman Optimal Assignment kernel generally produced better results than the Weisfeiler-Lehman Subtree kernel.

3 Dimensionality reduction

As described in the paper, we utilized Kernel Fischer Discriminant Analysis (KFDA) to perform dimensionality reduction. We then compared the effectiveness of KFDA to Kernel Principal Component Analysis (KPCA). Our results showed that KFDA outperformed KPCA in most cases, particularly in instances where the dataset was imbalanced and class information was significant. We experimented with various kernels and number of features, ultimately settling on an RBF kernel and reducing the number of features to 200.

4 Classification

After reducing the dimensionality of our dataset, we performed a classification task using Kernel-SVM with class-balancing and an RBF kernel. We then compared the performance of this approach with that of a RandomForest classifier. The results showed no

significant difference between the two classifiers in terms of classification accuracy.

5 Results

During our experimentation, we obtained positive results on our validation set, achieving up to 0.8 accuracy and 0.4 F1 score for the positive class. However, when evaluating the same model on our test set, the results were not as promising. To address this issue, we explored Subsampling to potentially improve our test set performance. Additionally, we attempted classification without reconstructing the set, and tried removing frequent subgraphs from the test set. While we made progress towards improving our model's performance, we acknowledge that there is still much to be done in terms of fine-tuning and refining our approach.

Project link

My Dropbox project can be found at https://www.dropbox.com/s/1bnzubvswcodgzm/KERNELPROJECTRGUIG.zip?dl=0[1]

References

[1] Tinghuai Ma, Wenye Shao, Yongsheng Hao, and Jie Cao. *Graph Classification Based on Graph Set Reconstruction and Graph Kernel Feature Reduction.* Neurocomputing, 2021.

[2] Lasse Regin Nielsen.Implementation available at *https://github.com/LasseRegin/gSpan*

[3] Xifeng Yan and Jiawei Han. gSpan: Graph-based substructure pattern mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 721–724, 2002

[4] Benyamin Ghojogh, Fakhri Karray, and Mark Crowley. *Fisher and Kernel Fisher Discriminant Analysis: Tutorial.* arXiv preprint arXiv:1906.09436, 2022.

[1] `https://www.dropbox.com/s/1bnzubvswcodgzm/KERNELPROJECTRGUIG.zip?dl=0`