Report_Q1

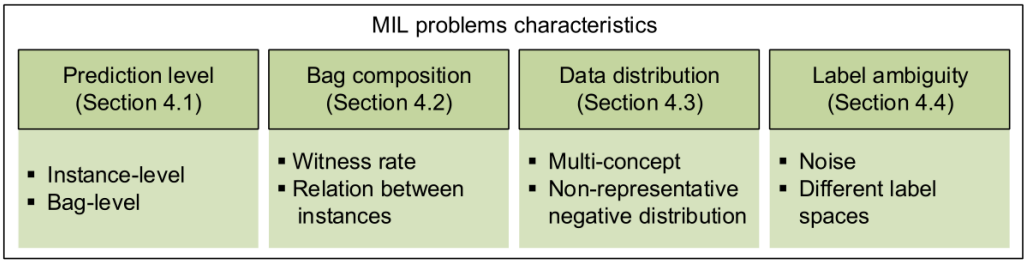
Report about the Deep Multiple Instance Learning Method used in a pan-cancer Study

Background

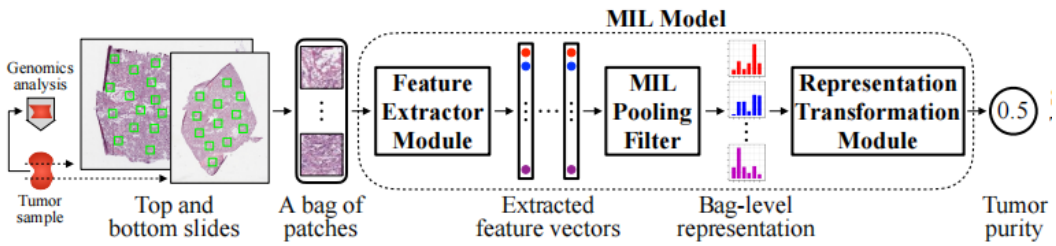
Tumor purity is the proportion of cancer cells in the tumor tissue. High-throughput genomic analysis is an important part of cancer research. However, the tumor purity of tumors greatly affects the collection and analysis of high-throughput data, so how to obtain accurate tumor purity is of great clinical significance. The genomics method or the pathologist's method is either insufficient in accuracy or requires a lot of human and time resources, so machine learning models came into being.

Algorithm (Multi-Instance Learning)

The MIL model was also originally proposed in the field of biomedicine by Dietterich et al.[1] in 1997. Most drugs are made up of small molecules that can bind to large proteins. For a molecule that can be made into a drug, one of its low-engergy shapes can bind tightly to the target site. A molecule can have multiple low-energy shapes, but at that time, scholars could only judge whether a molecule could be made into a drug, but not which low-energy shape of the molecule was working. Suppose we use a common classification algorithm and treat the low-energy shape of all drug-producing molecules as positive examples, and vice versa as negative examples. Then our training results will be very inaccurate because we have too many false positives. Based on this, Dietterich et al. [1] proposed the concept of MIL. He treats a low-engergy shape as an instance, and all the low-engergy shapes of a molecule form a bag. Molecules that can be used to make drugs are naturally positive bags, otherwise they are negative bags. The characteristics of MIL is shown in the image.[2]



In this paper, the author created a novel MIL model composed of three modules: feature extractor module, MIL pooling filter, and bag-level representation transformation module, shown as the following image.



Experiment

In order to implement the experiment with the MIL model on the MNIST dataset, I first downloaded the MNIST dataset and extracted the parts of digit 0 and digit 7 in the MNIST dataset. The next step is a key section, which is to package the MNIST data set into the bag required by the MIL model. In the bag composed of every 100 pictures, x pictures are digit 0, and 100-x pictures are digit 7, so my strategy is to generate a random integer k between 0-100, then put the first k data in the picture sequence of digit 0 in this bag, after that, put the first (100-k) data in the picture sequence of 7 in the bag. In this way, a total of 121 bags are obtained, and the proportion of 0 in each bag is also generated, which can be analogous to the tumor purity in the pan-cancer study, as the label of this bag.

Besides, the most important step is to find out the data type required by the model according to the source code, and then after slightly modifying the code of the result statistics, the experiment will be implemented successfully. But because of time constraints, I am trying hard to modify the data types required by the model, but not completely successful.

Conclusion

In general, although I did not successfully complete the experiment required in the stem, I tried my best to understand and do this experiment. In my previous study and research work, I have never been exposed to the MIL model. I think it is very inspiring for my future work after this experience, because one of the most important features of the MIL model

is that it does not require the data which most difficult to obtain in biomedical data, the mask labeled by pixel. Different from the pure classification networks which also don't need this data, the MIL network can obtain relatively more information. Besides, the data is packaged through slices, the efficiency of processing data is also higher than the general neural network model.

Through this experiment, I once again realized my insufficiency in data processing. Every time I begin to write a code of a different neural network, the part in which I face most problems is data processing. How can I convert the data into each forms I need when necessary?I still need further study and practice to solve this question.

The network conditions are also a major obstacle in this experiment. Due to limited network conditions, the git tasks often fail. Finally, in order to improve efficiency, I chose to upload files manually.

Reference

1. T.G. Dietterich, R.H. Lathrop, and T. Lozano–P´erez. Solving the multiple–instance problem with axis–parallel rectangles. Artificial Intelligence, vol.89, no.1—2, pp.31—71, 1997.
2. Carbonneau M A, Cheplygina V, Granger E, et al. Multiple instance learning: A survey of problem characteristics and applications[J]. Pattern Recognition, 2018, 77: 329–353.