DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Assignment 1

Finding applications of data mining

ENGG 5103: Techniques for data mining Deadline: 5:00 pm, Oct 7th, 2016 Submission: on elearning

Score counts: 4 points

Name: Lin Huangjing

SID: 1155072677

Problem 1

Problem: The MICCAI 2016 Challenge Organizers launched a challenge about how to automatic detect the potential breast cancer mitoses in the whole slide images (WSI). The goal is aimed at mining the very small potential mitosis cells from a very large-scale image (billion level pixels per image), and giving a grading report based on the counting of mitoses as diagnosis result. Generally, this procedure is time consuming clinically. It takes about 15~30 minutes for pathologist to diagnose one WSI. Also, heavy workload may lead to mis-diagnosis, when pathologists are tired. So how to mine the mitosis cells automatically is very meaningful to reduce the workload of pathologist.

Task: First, based on the breast cancer WSIs of patients, they want to annotate the potential mitosis cells in the WSIs. Second, they need to grade the WSIs into [not cancer, low risk, high risk] based on the number of mitosis cells in a certain field. Finally, select out the potential positive cancer cases.

Data:
1. Training data: Labeled breast cancer WSI cases annotated by professional pathologists. (Using the location of mitosis, to make training patches)
2. Testing data: WSI cases of other patients.

Method: Classification methods combined with fully convolutional neural network (FCN).

Actions:
1. Select a group of WSI cases as training samples that are annotated by professional pathologists.
2. Mitosis patches and negative patches should be extracted from the WSIs to build up the training dataset with preprocessing.

3. Train the classifier based on the preprocessed training dataset.
4. Using the well-trained classifier to traverse the WSIs to predict the potential mitoses.
5. Count the mitoses in a certain field of a WSI, and give an evaluation score.
6. Pick out the potential breast WSI cases for pathologists as reference.

My comments:
The method uses the WSIs of patients fetched in the clinic to predict the potential of the breast cancer. The classification techniques and algorithm efficiency plays a very important role here. Two points should be considered: 1. A suitable classifier helps boost the detection performance. 2. Image is very large (resolution is about 200000x100000), how to reduce the useless computation to speed up the algorithm should be considered as well.

Reference Link:
http://tupac.tue-image.nl/

Problem 2

Problem: The sinking of the TITANIC is one of the most famous shipwreck disasters in the 20th century. During her maiden voyage, the TITANIC sank after colliding with an iceberg, killing 1502 out of 2224 people. One of reasons that shipwreck led to such loss of life was that there were not enough lifeboats for passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class. This task is aimed at analyzing what sorts of people were likely to survive. Using data mining strategy to predict which passengers survived the tragedy.

Task: Based on the information of the part of passengers (like name, gender, age, ticket-class, etc.) and label (survive or not) to classify a new passenger into the category: 1. Survive or 2. Not survive. The information of other passengers are available.

Data:
1. Training data: the records of passengers with survival labels.
2. Testing data: the records of the remain part of passengers

Method: Classification method

Actions:
1. Select a group of people and mark down their survival labels.
2. Preprocessing the raw data with certain strategies.
3. Train a classifier based on the preprocessed data.

4. Apply this trained classifier to other passengers to predict their survival probability.

My comments:
In this task, preprocessing of the raw data is very important. How to transform and represent the raw for classifier would affect the final results significantly. Also a suitable classifier is important as well. The choice of classifier should be base on the scale of the dataset, neither too large nor too small. Too large classifier may lead to overfitting and degrade the performance. So this also should be considered.

Reference Link:
https://www.kaggle.com/c/titanic