Astros Mini-symposium

Friday, November 8, 2019

SR1 634

2:00 - 2:10pm:	Introduction from Astros R&D
2:10 - 2:30pm:	Bryan Florence
2:30 - 2:50pm:	Rachel Mills
2:50 - 3:10pm:	Saeed Sarmadi
3:10 - 3:20pm:	**Break**
3:20 - 3:40pm:	Aditi Singh
3:40 - 4:00pm:	Joey Xiaozhang
4:00 - 4:20pm:	Kazem Safari
4:20 - 4:30pm:	**Break**
4:30 - 4:50pm:	Vincent Poon, Ni Mei, Rainie Sun
4:50 - 5:10pm:	Rebecca Li
5:10 - 5:30pm:	Kate Nguyen
5:30 - 6:00pm:	**Networking**

Abstracts

Bryan Florence

2:10 - 2:30pm

The purpose of this project is to predict performance of MLB teams based on offensive stats using different machine learning algorithms. From Baseball-reference.com, a web scraping program was used to collect 24 offensive statistics from the years 2000-2018. Principal Component Analysis was implemented to analyze correlations between the statistics and to measure team performance. After reducing the dimensions of offensive statistics from 24 to 10, the Kmeans algorithm was used to place teams into one of two clusters; a cluster representing teams that made the post season, and a cluster of teams that did not make the post season. Kmeans successfully predicted the correct cluster with 66% accuracy for time period 2000-2018, although performance was higher in certain single seasons. For example, the algorithm performed at 80% accuracy for the 2018 season. K nearest neighbors was implemented to try and achieve better performance. For the years 2000-2018, the algorithm predicted the correct classification 83% of the time. It was also found that Hits, Runs, RBI, Doubles, Slugging, and Total Bases were consistent predictors of postseason appearances each year. Moving forward, I attempt to optimize the model by 1) including pitching stats in the model, 2) incorporating years as far back as 1973, 3) and compare the performance of the K nearest neighbors model with a model using a Nave Bayes Classifier.

Rachel Mills

2:30 - 2:50pm

My research mainly comprises of segmentation of complex data. The goal of my thesis is to improve patient response to immunotherapy in a reliable manner by implementing state of the art computer vision methods. Implementing instance segmentation on time-lapse videos will enable the progression of immunotherapy research. This is accomplished using an automated method to generate segmentations as labels and using a subsequent Mask R-CNN to perform instance segmentation. Preexisting aspects of this research includes cell detection, cell tracking, and cell death analytics performed over a large scale of time-lapse movies.

Additional work in segmentation include performing 3-d automatic detection and extraction of salt boundary tops within seismic data. Salt tops are essential for geologists within drilling companies to detect for the acquisition of oil. Therefore, performing detection and extraction in an automated way greatly increases the efficiency and accuracy of locating these essential structures. To perform reconstruction, automatic generation of labels is performed by thresholding synthetic seismic data, and 3-d semantic segmentation (3-d UNet) is used to train to detect salt tops in sparsely sampled data. As a result, this model is adaptable to predict salt tops in unseen seismic data.

Saeed Sarmadi

2:50 - 3:10pm

We aim to provide computational tools to aid the understanding of bacteria cells, and control and investigate their population dynamics. To do so, we need to provide tools for segmenting (delineating) and tracking the bacteria cells in two-dimensional movies of cell cultures. We propose a new method for segmentation and tracking of cells in a growing colony. We tackle three different problems. First image segmentation; this problem is delicate due to the low quality of the images and overlap of bacteria cells. Second, we have to track the movement of the bacteria cells across a time series of images (movies). Moreover, cells not only are displaced over time but can also split. We use different techniques, ranging from variational approaches to machine learning type algorithms, to tackle these challenges. We present some preliminary results on real and synthetic data.

10-Minute Break

Aditi Singh

3:20 - 3:40pm

Breast cancer is the most common cancer among women worldwide. The most widely used diagnostic method for breast cancer is visual inspection of histopathological images. Automating the classification of breast cancer histopathological samples into benign or malignant can make the diagnosis faster and more precise. In solving image classification problems with such high intra-class variability, Convolutional Neural Networks (CNN) have outperformed traditional machine learning approaches. However, they require more annotated data for training compared to conventional methods. Such a requirement creates a major obstacle towards using CNNs in the medical image domain. This paper explores active learning methods to train high-quality CNNs by annotating fewer but more informative data samples. We investigate two active learning approaches, based on entropy and Bayesian criteria, to classify histopathological tumor images into benign and malignant. Our approach yields a competitive accuracy by using only half of the training data compared to random selection approach. This finding makes active learning an appealing framework for building deep networks for biomedical applications where labeled data are often scarce.

Joey Xiaozhang

3:40 - 4:00pm

For NBA game data, the ordinary linear model takes the assumption of independence of the observations for free. In fact, games could be conducted for same match-ups in repetition or contain a team in common (e.g. game 1 - Houston vs LAL, game 2 - Houston vs LAC). Under these situations, the assumption of independence may fail. To analyze the potential correlations, the generalized estimating equations (GEE) model will take the correlations into consideration. I will use GEE model to analyze the NBA game data and discover their correlation structure.

Kazem Safari

4:00 - 4:20pm

Hyperspectral imagery (HSI) has emerged as a highly successful sensing modality for a variety of applications ranging from urban mapping to environmental monitoring and precision agriculture. Despite the effort by the scientific community, developing reliable algorithms of HSI classification remains a challenging problem especially for high-resolution HSI data where there is often a larger interclass variability combined with the usual scarcity of ground truth data. In recent years, deep neural networks have emerged as a promising strategy for problems of HSI classification where they have shown a remarkable potential for extracting joint spectral-spatial features efficiently. In this paper we propose a new strategy for HSI classification based on convolutional neural networks that leverages convolutional filters of various dimensions to handle spectral and spatial features. We show that our method achieves a very competitive classification result on the 2018 IEEE GRSS hyperspectral dataset a high-resolution dataset that includes 20 urban land-cover and land-use classes.

10-Minute Break

Vincent Poon, Ni Mei, Rainie Sun

4:30 - 4:50pm

With the rise of machine learning, and understanding of statistical learning, we can drastically improve our performance in many ways, including sports. Our focus are on human pose estimation and how to utilize the data it generated for us to help improve players performance and training new players.

Human pose estimation on its own is a sub field in computer vision, and its goal is to figure out where the person is in a picture or a video, and put a skeleton in the right place. That also means we will be able to collect information from the skeleton generated by the computer, and analyze it in a way that will provide use to the team.

In order to create something useful, we need to have a useful model to help us create meaningful connections between the data and the result we are looking for. For now, our goal is to try to estimate where a ball will be on a grid with the input as how the player throws it. More will be considered if the above model is successful.

Applying what we have learned in class, we opt to apply some classification algorithm with the body pose estimation data as input, and the pattern of the ball falling as the prediction. Using the historical game video as the training data, we hope to get the relationship between the body pose and the outcome of the actions. K-mean or CNN are considered algorithms. However, the interpretability of the model is the most important concern for us. So, we will balance the model accuracy and complexity to choose a proper algorithm or a hybrid algorithm to solve our problem.

The above is how we would approach the problem and we will now talk about some foreseeable challenges that we will have to overcome. First, data collection, in order to avoid costly data collection, we are considering using video online that is readily available to us, however, we are inexperienced with OpenPose.

The quality and quantity of the data might also be limited, as the information in a video is limited, eg we only knows did the ball get hit by the bat or not, but not how fast it was going. The number of examples we will be able to get our hands on are also limited before we can fully automate the data extraction process. This also the main limitation for us to use neural network as it often requires good sized dataset to learn.

Lastly, the result we generated can be put to use by the team. For example, help players to push the boundaries of their skills by reviewing the data to see how they can improve. Or formulate strategies by seeing if any particular way yields better results.

Rebecca Li

4:50 - 5:10pm

Conventional computational recovery is suffered from undesired artifacts such as oversmoothing, image size limitations and high computational cost. The use of deep generative network (GAN) models offers a very promising alternative approach for inexpensive seismic data acquisition, which improved quality and revealing finer details when compared to conventional approaches or pixel-wise deep learning models. We are one of the pioneers to apply a pixel inpainting GAN on large, real seismic compressed image recovery.

Kate Nguyen

5:10 - 5:30pm

Natural environments change over many different timescales. To make the best decisions organisms must flexibly accumulate information, accounting for what is relevant, and ignoring what is not. However, many decision-making studies focus on sequences of independent trials in which the evidence gathered to make a choice and the resulting actions are irrelevant to future decisions. An important and naturalistic aspect of such experiments is a probabilistic schedule of rewards, in which sometimes correct responses are not rewards and sometimes incorrect responses are. For example, birds may correctly identify a fruitful tree but have difficulty harvesting the fruit. Normative theory has been developed for such tasks when reward is the sole evidence (e.g., two armed bandit tasks), but less is known about how ideal observers integrate probabilistic rewards interspersed with noisy evidence of the more often rewarding choice. To understand decision-making under more natural conditions, we propose and analyze models of ideal observers who accumulate evidence to freely make choices across a sequence of correlated trials and receive uncertain feedback.

30-Minute Networking