Peter Hotvedt

MP 574

Sean Fain and Diego Hernando

July 31, 2020

Final Project Report

For the final project in MedPhys 574 , my partner Noah Schweitzer and I opted to investigate adaptive k-means clustering and its relationship to my research in Dr. Fain's lab. In my work in Dr. Fain's lab, I have been investigating the relationship between Ventilation Defect Percentage(VDP) and various metrics like Parametric response Mapping and mucus plugs in the lungs. The VDP values have come from Hyperpolarized Helium-3 Gas (HP ^{3}He) MRI images that have been segmented using an adaptive k-means algorithm though most, if not all, of this work was done before I began my work on the project. Therefore, I thought it would be interesting to see if I could replicate the results that had been derived by my coworkers and course instructor. The relationship between the project and the course is very straightforward; the algorithm being investigated is one that segments a lung image, and this kind of process is one of the many machine-learning process taught in the later portions of the MP574 course. Furthermore, while k-means algorithms are covered, the variations of the algorithm are only briefly mentioned, hence the study of adaptive k-means clustering is an expansion of the knowledge developed in class. The adaptive k-means clustering algorithm is a good fit for the clustering of ventilation defect percentage in this case because it is able to take the lung images gathered from the HP ^{3}He MRI and visualize how healthy a pair of lungs may be and where there may be damage. Additionally, given the replicative nature of this project with work that has already been done, it would be orthogonal to the project to try and develop a different algorithm to produce the same results.

The final project plan required two basic deliverable items: a paper explaining the mathematics and context of the problem and an algorithm with images and VDP calculations derived from the group-created script. The first deliverable, the paper, focused on the mathematical aspect of the project as it went into the details of how the algorithm operates. In order to run a "naive" k-means algorithm, two basic mathematical steps are needed, a sum-of-squares step and an centroid updating step. Both steps will be explained in the mathematic sense, with a specific focus on the mechanics behind the math of the algorithm

To begin to understand the k-means algorithm in general, the mechanics behind the idea of minimizing a distance in a space. The effective distance measurement is a simple n-dimensional least squares problem that find the best solution to $Ax = b$ by computing the Euclidean Norm on the difference $||Ax - b||^2$ and minimizing it. For the context of this problem, the $|| \cdot ||$ is just the 2-norm which is defined as:

$$|| \cdot || = \left(\sum_{i=1}^{n} x^2 \right)^{1/2} \tag{1}$$

and is to be assumed as such unless specified otherwise. The idea is to find some solution vector for x^* that minimizes the 2-norm such that $||Ax^* - b||^2 \leq ||Ax - b||^2$. It should be noted that A and b are meant to be taken in a general sense, and that they vary for different least-squares problems. In this case, Ax_i can be written as X_i and is an observational data point in a given cluster, and b_i is written as the centroid of a cluster μ_i This is done across the set of all the data points within the general data set S (where S has k partitions or clusterings) in order to minimize the intra-cluster variance. To summarize the math, the problem is to essentially minimize equation (2) with respect to each individual clustering in the data space.

$$\sum_{i=1}^{k} \sum_{x \in S_i} ||X_i - \mu_i||^2 \tag{2}$$

A time-consuming, yet functional, version of the algorithm involves two basic steps: an assignment step and an updating step. In the assignment step, each clustering is assigned to the nearest mean value using the least-squares solution to the 2-norm for the mean at an iteration (t). This is mathematically similar to the process previously outlined, but done explicitly in equation (3)

$$S_i^t = \left\{ X_m \ : ||X_m - \mu_i^t||^2 \leq ||X_m - \mu_j^t||^2 \quad \forall j, 1 \leq j \leq k \right\} \tag{3}$$

In this case, X_m is the general observation in the clustering, and is only assigned to one cluster S_i. This work is done such that despite that fact that X_m might theoretically belong to more than one clustering, there are no data observations in more than a single clustering. The updating step then takes all of the observations X_j in a given clustering S_i and finds the new average location of the clustering, therein updating it appropriately. This is a relatively simple mathematical operation as it is just determining the average location of the cluster, and assigning it to be the new center.

$$\mu_i^{t+1} = \frac{1}{|S_i^t|} \sum_{X_j \in S_i^t} X_j \tag{4}$$

From this averaging, the algorithm continues until the difference between the previous cluster centroids and the updated centroids are negligible, or $\mu_i^{t+1} - \mu_i^t \approx 0$. The tolerance on this is adjustable as needed, though the important part is to verify that the algorithm terminates. This is the general sense of the algorithm and was determined and developed for the context of this problem.

This algorithm was written into a MATLAB code and applied to fifteen specific cases. These lung images were provided by Katie Carey, a member of Dr. Fain's lab, and were specifically chosen to show a variety of differeing level of VDP. Of the fifteen cases, one was a Healthy patient, seven were Mild/Moderate cases, and seven were Severe cases. In order to prevent result bias, all but three of the cases were blinded by Katie, only being revealed after our MATLAB script was written properly and our results were relatively close to the existing VDP results. There was not much post-processing that needed to be done, though much of the progress depended on being effective communicators with Katie and making sure that we had consistent access to the data and files we needed. Katie was very helpful in this regard and I am very thankful for all of her help with this project.

Examples of the outputs of the image segmentation are found below for a Healthy, Mild/Moderate, and Severe case. In each figure, the image segmented by our script and the "Gold Standard" algorithm from the previously done work are shown.

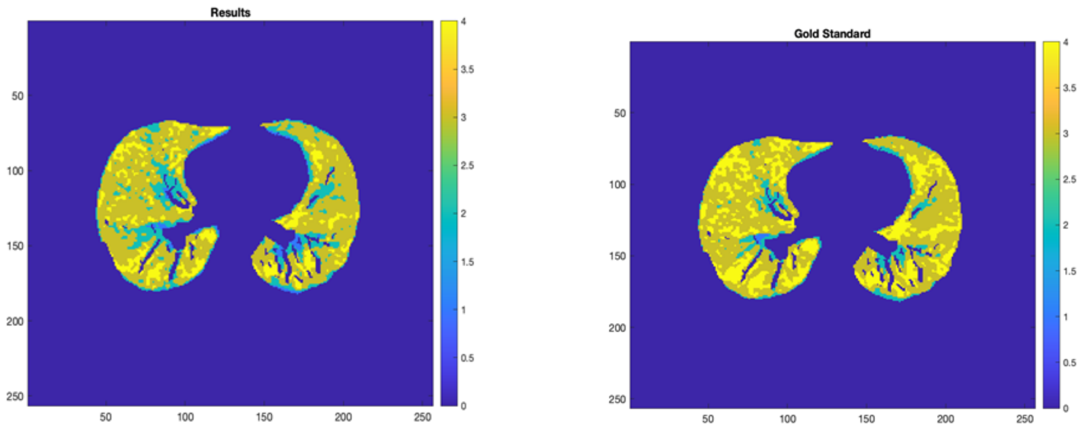

Figure 1: The clustering from our script and the gold standard for the Healthy patient

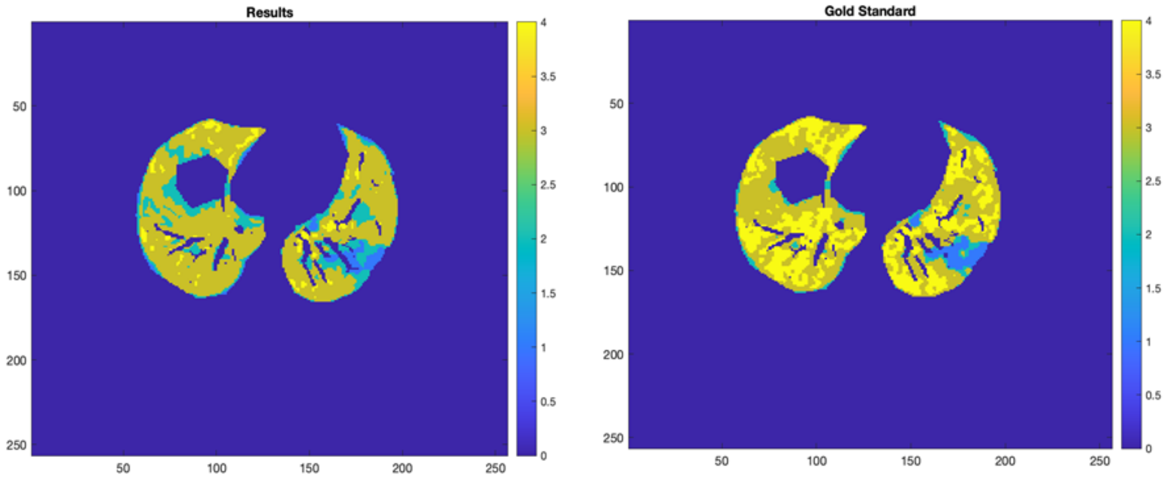

Figure 2: The clustering from our script and the gold standard for the Healthy patient

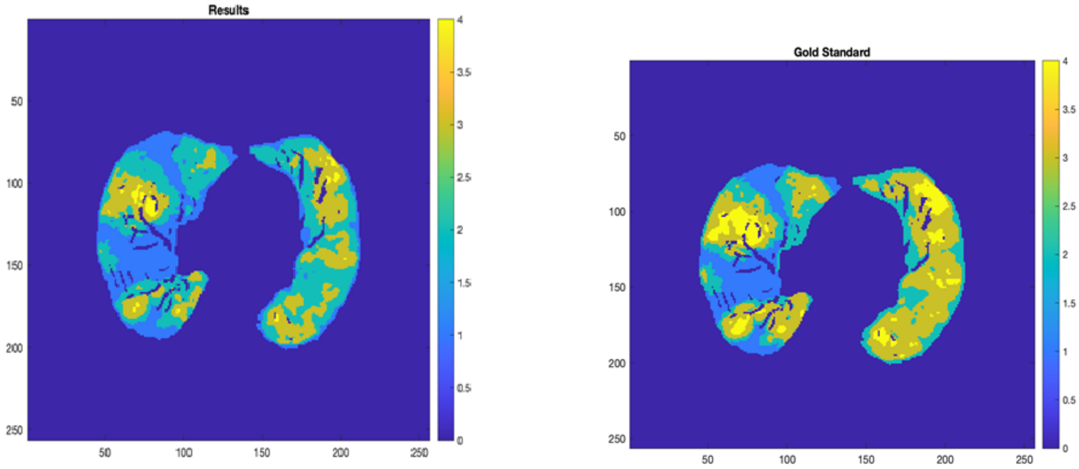

Figure 3: The clustering from our script and the gold standard for the Healthy patient

In each of the clusterings, while not perfect, there is a pretty close match of where there is true ventilation defect as designated by a 1 on the colorbar. There are, however, some discrepancies in the healthy regions of the lung as seen most visibly in the severe case. The differences in clusterings between cases where pixels were labelled 1 and 2 is very low, while, in contrast, pixels clustered into 2 and 3 vary notably. This is most likely the case because of the focus on clustering the severe defect and could be solved by repeating the clustering process while masking out any pixels clustered in the VDP group. However, the healthier parts of the lung are not necessarily the focus of the paper, so they can more or less be ignored when calculating VDP.

Our results from the algorithm turned out to quite similar to the previously found results, as shown in the table below:

Subject ID	Severity	VDP (Katie) [%]	VDP (Us) [%]
80-811-151	Healthy	0.13	0.13
80-811-023	Mild/Moderate	2.41	2.84
80-811-027	Mild/Moderate	1.27	0.62
80-811-131	Mild/Moderate	2.84	3.52
80-811-135	Mild/Moderate	2.15	1.92
80-811-139	Mild/Moderate	0.26	1.44
80-811-145	Mild/Moderate	5.31	5.01
80-811-118	Mild/Moderate	3.09	3.48
80-811-071	Severe	31.66	38.43
80-811-011	Severe	25.9	22.35
80-811-110	Severe	8.03	12.45
80-811-070	Severe	7.68	12.84
80-811-059	Severe	5.71	8.85
80-811-074	Severe	21.4	23.15
80-811-057	Severe	9.66	17.57

Table 1: The comparison of our VDPs vs the Gold Standard based on patient data

While the results seem to match Katie's data quite well, there is that case that for a majority of the severe cases and some of the Mild/Moderate cases, our algorithm yields a much higher VDP. This may be due to the fact that our algorithm overclustered a little bit, therein making the data biased towards a larger VDP. However, aside from this, the data seems to match quite well. We also wanted to verify repeatability of this algorithm and see if the VDP would change over the course of running multiple trials. Sure enough, the algorithm produced essentially identical results after many (5) trials. This is a reassuring thing as it shows the algorithm will give the same results over many repetitions and that other using our algorithm would also get our VDPs. Ultimately it is difficult for us and our methods to yield the exact results that Katie got without repeated testing over a longer span of time, but it is likely that with a few tweaks to the code to remove a majority of the over-clustering, our results would approach the “Gold Standard”. The code does, however, approximate the areas of ventilation defect very well, meaning that our work does seem to be a very good initial representation of the actual situation.

All in all, this project went quite well for us being a two person group. Noah and I worked very well together, a continuing trend since we have had nearly every class together since Fall 2018. There was a very strong group dynamic as we were able to communicate what needed to be done with each other quite well. It also helped significantly that I had a very strong background with the entire research process and could easily define terms and case numbers for Noah when the situation arose. One of the important things to note is that we verified each other’s work as the code and papers were being written. This removed some of the challenges that we may have had faced had we not communicated as effectively as we did. The most difficult part of the entire project was making sure we had the data at the right points in time, but that issue was quickly remedied by Katie’s helpfulness.

In addition to MP574, I have been taking other courses related to the deeper studies of linear algebra, including a Numerical Linear Algebra course (CS513). This course has run in parallel with this course with respect to determining algorithms using least-squares problem solving methods and data-fitting algorithms, so my learning in MP574 has been even further supplemented to where I felt comfortable attempting to develop this algorithm for the assignment. The difference between CS513 and this project is that I was able to combine and apply skills I learned in both classes and apply them here, explaining both theory and the practice all in one fell swoop. On a personal level, I feel like I really developed a lot of strong computational and numerical analysis skills by having that parallel overlap that both courses provided.

I also learned quite a bit about the previous work that had been done by those who have been working on the Image-Guided Bronchoscopy and Co-localization project. I have been working on the stages involving the CT images but I began work after the MRI images had already been segmented. Earlier on in my research career, I was still somewhat confused as to how the MRI images we used had been portioned in such a way that we knew how they VDP was developed. I did my own research during the process to get an idea of what segmentation was, but it was not until this process that I really had a firm understanding of the methodology that this project used. Without sounding too cheesy, it was really reassuring for me to get values very close to the ones that had been previously found and it made me realize that I am perhaps more capable than I thoughts. I developed a sense of appreciatino of the work that went into developing the workflow so that I could continue the PRM and tPRM analyses that we are now working on.

There are some obvious limitations that we faced while conducting this study. The most obvious case is that given the nature of the COVID-19 pandemic, it was often very difficult being able to work on the project without constantly communicating. In a regular setting, work on this project may have been accelerated due to the nature of meeting face to face yielding a lot more productivity compared to texting ideas. This would have also likely increased our ability to get our data, as a majority of the time we had to wait for an email from Katie. Of course, this limitation did not end up impacting the quality of our actual results that much, just the amount of time we had in terms of getting the results. Given more time, it is likely we could have done this segmentation for the majority of the SARP III patients.

As previously noted, a lot of work with this segmentation is already underway in Dr. Fain's lab. Presently, the VDP gathered from this segmentation process is being compared with CT images in order to compare the topological measures of the lung and how the lung responds to potential air-trapping or other lung disease. Additionally, this VDP is being measured against the parametric response mapping with respect to the air being expelled from the lungs in spirometry tests in order to determine potential correlations. Furthermore, much of the data is being compared with mucus scoring as well, as it would be very illuminating to discover if a defect is the result of a mucus plug in the lungs, or if there is not a relationship whatsoever. There are plenty of other uses that this segmentation could be used for, but this is just what is presently going on.

One possibility of using this segmentation process could be to observe eventual HP ^{3} He MRIs of COVID-19 patients and see if there are any remaining ventilation defects as a result of contracting the disease. The segmentation process outlined in the report is a fairly general method, so it can easily be applied to other measures as well in attempts to verify if there are any defects in COVID-19 patients, or if patients with high VDP are more susceptible to the adverse effects of COVID-19. Given the pandemic, research into the virus is a very hot topic, and this image segmentation process could easily be used in that setting. This brainstorming process is very blue-sky-ish and there would likely be many other factors to consider, but further research with this segmentation algorithm is certainly feasible.

Ultimately, this final project went very smoothly and I am very pleased with our results. The learning process and the results we got were both quite good and I am very pleased with the work that went into this project. While there may have been some general over-clustering in our results, in general, we were able to effectively get results that were very close to the previously found solutions. In general, the project was a great way to practice and fully understand the ways in which image segmentation takes place, and was a great way to visualize a lot of the work that took place over the course of the semester.