Research Statement

Manoj THULASIDAS

School of Computing and Information Systems,
Singapore Management University

Tel: (65) 6828-9625; Email: manojt@smu.edu.sg

30th December 2020

Background

For a faculty in the education-track, the research focus is expected to be pedagogical in nature. It includes tools and techniques that improve the quality of our teaching. It may also include pedagogical topics that illustrate basic concepts. Based on this observation, I have developed a two-pronged research strategy going forward.

Pedagogical Topics: Developing algorithms and investigating their statistical underpinnings is in line with my quantitative and scientific background. A large part of my initial research work at SMU was, therefore, in developing algorithms appropriate for teaching the fundamentals of data science and statistical concepts, which resulted in a paper ("Nearest Centroid: A Bridge between Statistics and Machine Learning") at TALE 2020.

Pedagogical Tools and Techniques: In the second term of 2019, the COVID situation made us rethink on the final examination modality for COR1305, which required the students to upload an Excel file for the instructors to evaluate. Since the students would be working from home, the integrity of their work had to be preserved. Taking this challenge on, I embarked on a solo project to ensure the security of the answer books. Going beyond the challenge, I also automated the grading process, which was adopted by other faculty members in AY2020-21 term 1. This work was presented as a paper ("Secure Answer Book and Automatic Grading") at TALE 2020.

For 2021 and beyond, I plan to expand on these two prongs of my research focus, while keeping in mind that some of the foundational algorithms developed as pedagogical topics may qualify to be included in the mainstream research in data analytics. One such topic is the so-called K Selection problem (determining the right number of clusters to form in K-Means clustering), which could be made quantitative with the help of some basic statistical principles. I have been working on this topic for the last three years, and it is probably ready for publication.

Research Plan

Pedagogical Topics

1. The K-Selection Problem and its Solution (TALE)

Traditionally, the solution to the K Selection problem (of determining the right number of clusters to form in K-Means clustering) taught in our classes is the "elbow" method, where we look for a kink in the evolution of a quality metric (typically SSE, the sum of squared errors). While exploring potential metrics for clustering quality, I came across some another index, well-established in the literature, which works significantly better. In this work, I plan to compare the elbow method to some of the indexes available including Variance Ratio Criterion (VRC) and publish it as a recommended topic for introductory data analytics courses.

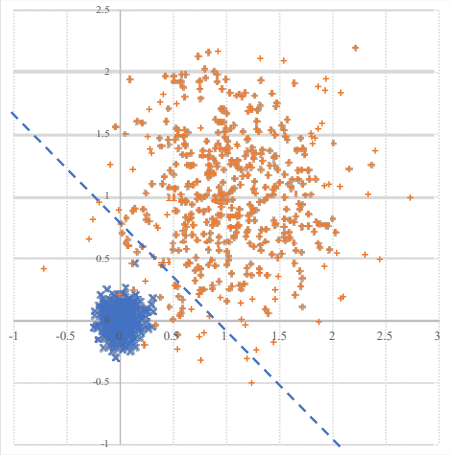
2. Topic Assignment to Sentence Tokens Using Nearest Centroid (COLING)

In text analytics, we teach LDA as a topic modelling algorithm, which generates a probabilistic topic model of the corpus under study. It then models each document as a mixture of topics, without assigning

topics to any specific part of the document. The idea behind this project is to map topics to sentence tokens (or some other logical sub-segment) within documents using classifiers and establish its validity.

3. Recursive K-Means or SVD for Clustering

One of the known drawbacks of K-Means algorithm is that it expects uniformity in terms of cluster membership and variance. If the variances of clusters are very different from each other as shown in Figure below, K-Means clustering may fail, which can be mitigated if we could normalize the data using the variance of the corresponding cluster. We cannot do it, however, because we do not have the cluster affiliation before the clustering is performed, setting up the recursive problem. This research project proposes to break the recursion using the algorithm listed in below.



Input: Data set
Output: Cluster Assignments
Initialization:
1. Normalize the data set
2. Select variables and number of clusters
   LOOP Process
3. while Not Converged do
   a. Perform K-Means clustering
   b. Segment the data using the clusters
   c. Normalize each segment individually
   d. Combine the segments
4. end while
5. return Last K-Means results

The figure shows two clusters with different variances. The blue tighter cluster scavenges some members from the amber cluster because the cluster boundary line is the bisector, equidistant from the two centroids. The textbox describes the recursive algorithm to solve the problem of unequal variances. It nullifies the effect the differences in variances by normalizing away the measured standard deviations. This project is expected to use simulated data to illustrate the algorithm and real data for proof of viability.

4. A Clustering Quality Metric Based on Centroid Locations (Submitted to ICDM 2021)

While teaching K-Means clustering to my Analytics Foundation students, I hit upon the notion of using the significance of the distances between the cluster centroids to help select the "best" variables to use. I was illustrating why some variables in the Iris data set were better than others in partitioning it to the three species. The distance between the "better" variables, when segmented by the species names, have a higher significance (in terms of z-scores) compared to the rest.

In a real clustering problem, we do not have class labels such as species names. I tried to explain to my students that we could then use the clustering results of an initial run as a proxy to the actual class labels and compute the z-scores. This concept was a bit too abstract for the undergraduate students and I dropped it from my subsequent teaching of the course. However, I kept the idea and later developed it into a metric to quantify the clustering quality.

The first paper based on this work is "A Quality Metric for K-Means Clustering" (Presented in ICNC-FSKD 2018). This paper addressed the typical practical questions before performing K-Means clustering on any data set:

1) How many clusters should be form? What is the right value of K?
2) What variables should we use?

The first paper, although well-received at the conference, was rather limited in its scope. Among other things, it focused solely on the properties of the centroid locations. In a significant extension to the original work, I added a whole host of improvements in this second paper ("Quality Metrics for K-Means Clustering Based on Centroid Locations") and submitted to ECAI 2020, where it received positive reviews at the author rebuttal stage, but was not eventually accepted. Based on the reviews at ECAI, I have improved the paper, especially by adding synthetic data, and submitted it to ICDM 2021.

Pedagogical Tools and Techniques

1. Tracking of Learning Objectives and Competencies using Online Quizzes: A Case Study (TALE)

One advantage of running continuous or summative assessments as online quizzes is that it generates a wealth of data. We can then analyze the data, along with additional information, to track student learning and generate insights for improving pedagogy. In this case-study style paper, I plan to the statistics on the continuous assessment and the final exam quizzes from our eLearn platform, decorate it with learning objectives and competencies of two courses (COR 1304 and IS450) and analyze it. The aim is to quantify the coverage of the assessments and the efficacy of teaching (from the student performance per learning objective and competency). The results should enable us to improve our teaching of the course in the future.

2. Randomized Cold Calls (TALE)

Cold Calls are an effective pedagogical technique to keep students engaged in the class activities. One common problem, however, is to ensure true randomization, especially if class participation is rewarded in grading. If not randomized, the instructor may unwittingly call on students whom he recognizes or who have participated previously, skewing the participation toward the more vocal of the cohort. Using a simple program, but cleverly integrated with the presentation software, we can improve the student engagement level. A straight-forward randomization of the names called, however, would not work, because of the so-called birthday paradox. This paper will provide a survey of the existing literature on cold calls and describe the method I adopted to implement it along with the lessons learned.

3. Moderation Tool (TALE)

Most universities have prescribed grade distributions for the final letter grades in every class or cohort. These distributions are achieved by deliberate manipulations of the component grades or by applying moderation. Instructors end up spending a large amount of time on it, often making the process unduly subjective. However, at the heart of it, the distributional requirement can be boiled down to a couple of simple principles: (1) A student with a lower score than another student should never get a higher grade than the latter, and (2) Ideally, the grade boundaries (the cut-off scores between letter grades) should have relatively large gap. I have developed a robust tool implementing just these principles and will explore whether it can be shared at a future TALE conference as a productivity tool.

4. From Discussions to Predictions

During pre-COVID days, I had my students in the last term participating in the Piazza discussion forums as part of out-of-class activities. Based on their participation statistics, I used to award up to 5% of their final grade. Looking at text data of the discussions on the forum, I feel this dataset may contain more information about the student performance. Can we predict student grade based on the participation statistics? Can we identify the students needing help using the predictions? Can we sharpen the findings using text-analytics? These questions formed the basis of the new pedagogical research project I had planned to work on in 2020, but it was disrupted by the pandemic. I will look into it again in 2021, in collaboration with some other faculty members.